Enhancing Communication Accessibility: A Deep Learning Approach for Assistive Hand Gesture Recognition in Speech Disability Communities

Zahariah Manap[*1], Abdul Haiqal Baharin[2], Suraya Zainuddin[3], Juwita Mohd Sultan[4], Azita Laily Yusof[5]

[1,2,3,4]Centre for Telecommunication Research and Innovation (CeTRI), Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka, Malaysia

[5]School of Electrical Engineering College of Engineering, Universiti Teknologi MARA, Malaysia

[*]Corresponding Author

DOI: https://dx.doi.org/10.47772/IJRISS.2025.909000778

Received: 27 September 2025; Accepted: 03 October 2025; Published: 30 October 2025

ABSTRACT

Individuals with speech disabilities face significant barriers in daily communication, leading to social isolation. While augmentative and alternative communication (AAC) devices exist, they often lack intuitiveness and real-time performance. This study investigates the efficacy of deep learning-based object detection models to create a non-intrusive, vision-based hand gesture translator. A custom dataset of 275 images representing five essential gestures ("Hello," "Thank You," "Yes," "No," and "I Love You") was constructed and used to train three state-of-the-art architectures: SSD MobileNet V2 FPNLite 320x320, SSD ResNet50 V1 FPN 640x640 (640x640), and EfficientDet D0 512x512. Performance was evaluated based on precision, recall, and loss metrics. The SSD MobileNet V2 model demonstrated superior performance with a precision of 0.8869 and a recall of 0.8867, offering an optimal balance between accuracy and computational efficiency. In subsequent real-time prediction tests on 100 samples, the system achieved an overall accuracy of 95.5%. The results underscore the potential of lightweight deep learning models in developing affordable and efficient assistive technologies, providing a robust foundation for enhancing communication accessibility and social inclusion.

Keywords: Assistive Technology, Speech Disabilities, Augmentative and Alternative Communication, Hand Gesture Recognition, Real-Time Translation

INTRODUCTION

Communication is a fundamental pillar of human interaction, essential for expressing needs, building relationships, and participating in society. However, for individuals with speech disabilities, a condition that can stem from neurological disorders like cerebral palsy or Amyotrophic Lateral Sclerosis (ALS), traumatic injury, or post-operative recovery, this fundamental ability is severely hindered. This barrier often leads to profound social isolation, frustration, and a diminished quality of life [1], [2].

Traditional Augmentative and Alternative Communication (AAC) devices, such as speech-generating tablets or picture boards, provide crucial support but are frequently characterized by high cost, steep learning curves, and a lack of natural, intuitive interaction [3]. The recent convergence of advanced machine learning and computer vision offers a transformative opportunity to overcome these limitations. Vision-based hand gesture recognition systems present a promising alternative, leveraging non-intrusive cameras to translate natural gestures into text or speech in real-time [4].

While previous research has demonstrated the feasibility of machine learning for gesture recognition [5], many existing systems grapple with challenges in robustness, accuracy under varied real-world conditions (e.g., lighting, angles), and computational efficiency necessary for real-time application on accessible hardware.

This work aims to address these gaps by conducting a rigorous comparative analysis of modern, efficient deep learning architectures for the specific task of translating a core set of communicative gestures. The primary objectives are:

1. To develop and annotate a custom dataset of five fundamental hand gestures captured from diverse perspectives to ensure model robustness.
2. To train and evaluate three leading object detection models: SSD MobileNet V2, SSD ResNet50, and EfficientDet D0, using standardized metrics of precision, recall, and loss.
3. To validate the performance of the optimal model through real-time prediction tests, assessing its practical applicability as an assistive communication tool.

The findings of this work contribute to the growing field of assistive technology by identifying an optimal model that balances high accuracy with low computational demand, thereby paving the way for the development of cost-effective and user-friendly communication aids.

LITERATURE REVIEW

The field of assistive technology for communication has evolved significantly, from simple picture boards to sophisticated electronic devices. Modern AAC devices, while powerful, often rely on screen-based navigation or complex encoding systems that can be cognitively demanding and slow, hindering fluid conversation [3].

Augmentative and Alternative Communication (AAC)

Hand gesture recognition exists within the broader field of AAC, which encompasses a wide spectrum of techniques and technologies designed to support individuals with complex communication needs [5]. For people with speech disorders or motor impairments, such as those resulting from stroke, cerebral palsy, autism, or genetic syndromes, AAC serves as a critical tool for expression and social participation. This spectrum ranges from unaided systems such as manual signs, gestures, and facial expressions to aided systems that involve external tools [5], [6].

Aided AAC itself varies from low-tech, tangible methods like the Picture Exchange Communication System (PECS) and communication boards, to high-tech, sophisticated devices. This includes Speech Generating Devices (SGDs) and systems utilizing eye-tracking technology [6]. A significant recent advancement in high-tech AAC is the development of Visual Scene Displays (VSDs), which are interactive interfaces that use images or videos of real-life events to provide a more intuitive and contextual framework for communication, particularly for beginner communicators [6].

The ultimate goal of these diverse AAC solutions is to help users develop linguistic, operational, social, and strategic competencies [6]. However, many existing high-tech solutions face limitations in cost, portability, and real-time performance. They can be prohibitively expensive, physically cumbersome, and often lack the intuitive, naturalistic interaction that unaided communication provides.

Framework for Vision-Based Hand Gesture Recognition

The integration of computer vision into AAC devices, particularly for hand gesture recognition, typically follows a standardized processing framework. This framework, conceptualized in Fig. 1 [7], begins with the acquisition of visual data via an input device such as a standard RGB camera or depth sensor. The raw data of the captured gesture then undergoes image processing stages, which often include preprocessing (e.g., noise reduction, contrast adjustment), segmentation to isolate the hand from the background, and feature extraction to identify key points, contours, or shapes.

These processed features are then passed to a core gesture classification module, which is increasingly powered by deep learning models. This module maps the input features to a predefined gesture label. Finally, the output of this process is the recognized gesture, which can be translated into text or speech as part of an AAC system.

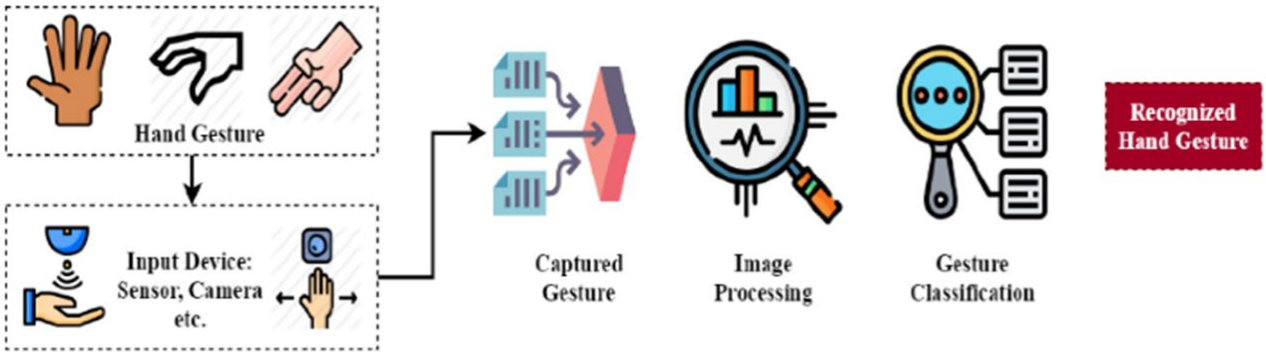

Fig. 1 Conceptual framework for vision-based hand gesture recognition systems [7]

Deep Learning for Computer Vision in AAC

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized computer vision tasks, including gesture recognition. Early approaches relied on traditional image processing techniques like Haar cascades or Histogram of Oriented Gradients (HOG) paired with classifiers like Support Vector Machines (SVM) [8]. These methods, while foundational, were often brittle, requiring controlled environments and struggling with variability in lighting, background, and user appearance.

The shift to deep learning models, such as the Single Shot Multibox Detector (SSD) [9] and You Only Look Once (YOLO) [10] architectures, enabled end-to-end learning of features directly from pixels, dramatically improving robustness and accuracy. Models like MobileNet [11] were specifically engineered for efficiency, utilizing depthwise separable convolutions to reduce computational cost and model size, making real-time inference on mobile processors feasible. This is a critical advancement for assistive technology, where affordability and portability are paramount.

Recent studies have successfully applied these models to sign language recognition [12], [13]. This research direction represents a convergence of AAC principles and advanced computer vision, aiming to create digital systems that are as intuitive as unaided gesture-based communication but with the transformative power of technology. However, a gap remains in developing lightweight systems focused on a core set of communicative gestures, rather than a full sign language lexicon, that can operate in real-time on low-cost hardware. This study seeks to bridge the gap between the high cost of specialized AAC devices and the need for natural, accessible communication. It does so by leveraging efficient deep learning architectures to create a vision-based AAC tool that is both affordable and intuitive, directly addressing the limitations of current high-tech solutions identified in the AAC spectrum.

METHODOLOGY

The research methodology was structured into three sequential phases: data collection and preparation, model training, and performance evaluation, ensuring a systematic approach to developing and validating the gesture translation system.

Data Collection and Dataset Preparation

A foundational custom dataset was created to train and evaluate the models. It comprised 275 RGB images capturing five distinct gesture classes: “Hello”, “Thank You”, “Yes”, “No”, and “I Love You” (see Fig. 2). Images were acquired using an HP True Vision FHD webcam at a resolution of 640×480 pixels. To emulate real-world usage and ensure model generalization, samples for each gesture were captured from multiple viewpoints, including frontal, lateral (left/right), and tilted (up/down) angles, introducing natural variations in scale and

orientation. While the sample size of 275 images is a starting point for this proof-of-concept, it provides a solid basis for the comparative model analysis which is the focus of this work.

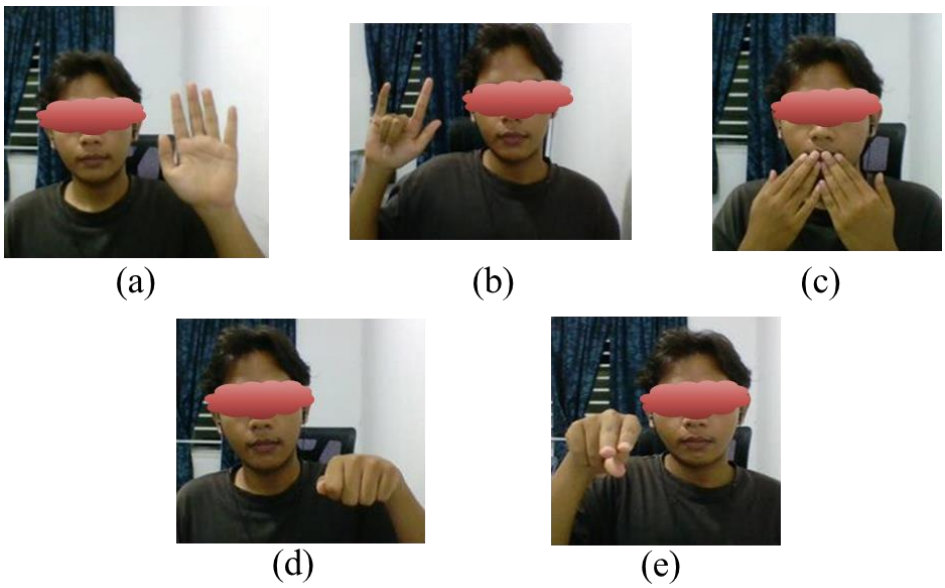

Fig. 2 The five gesture classes: (a) Hello, (b) I Love You, (c) Thank You, (d) Yes, (e) No

Each image was meticulously annotated using the LabelImg tool to generate bounding boxes around the hand region. The annotation process, demonstrated for the "Hello" gesture in Fig. 3(a), produced XML files containing the normalized coordinates of each bounding box (see Fig. 3(b)). This manual process ensured precise localization while inherently incorporating variability in hand size and position, which benefits model robustness.

The use of normalized coordinates for the bounding boxes is a critical preprocessing step. It ensures that the model learns spatial relationships independent of the original image resolution, which is vital as different models require different input sizes during training and inference (e.g., 320x320 vs. 640x640).

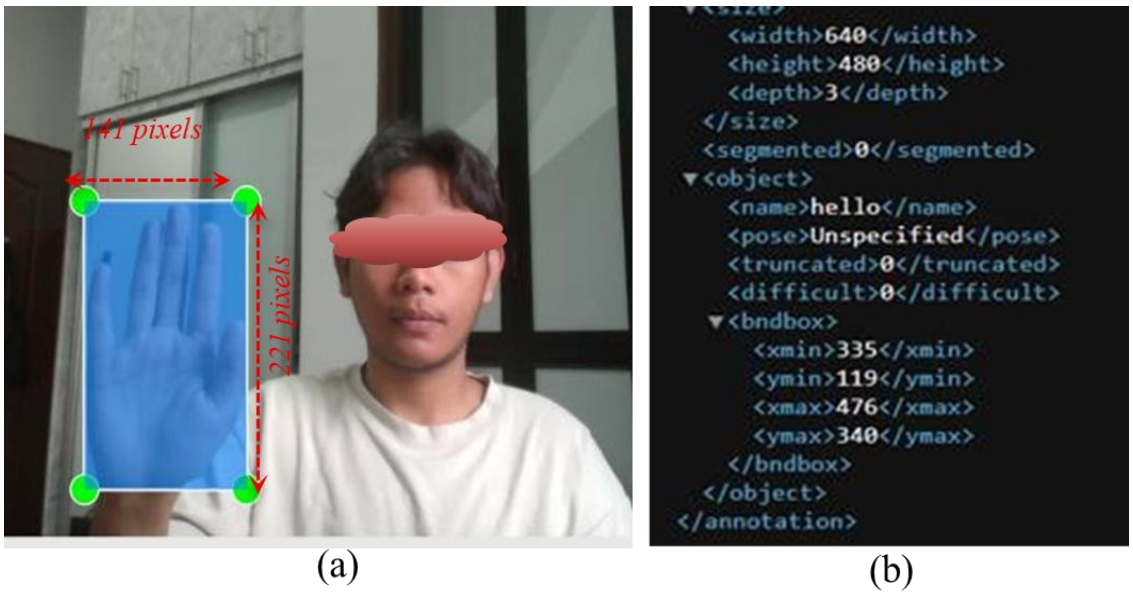

Fig. 3 Annotation process for the 'Hello' gesture: (a) Bounding box drawn in LabelImg, (b) Excerpt of the resulting XML output

The finalized dataset was partitioned into an 80:20 split, resulting in 220 images for training and 55 held-out images for testing, ensuring the evaluation was performed on completely unseen data.

Model Selection and Training

Three object detection models were selected for comparative analysis, each representing a different trade-off between speed and accuracy:

1. SSD MobileNet V2 FPNLite 320x320: A lightweight architecture optimized for fast mobile and embedded vision applications. Its use of depth wise separable convolutions and a feature pyramid network (FPN) neck makes it efficient for detecting objects at multiple scales without a heavy computational burden.
2. SSD ResNet50 V1 FPN 640x640 (640x640): A deeper architecture offering higher potential accuracy at a greater computational cost. The ResNet50 backbone, with its residual connections, enables the training of much deeper networks, potentially capturing more complex features. The higher input resolution (640x640) can preserve finer details.
3. EfficientDet D0 512x512: A modern family of models that scales network depth, width, and resolution efficiently. It represents a state-of-the-art balance between these factors.

All models were implemented using the TensorFlow Object Detection API. To ensure a fair comparison, each model was trained for 6,000 steps with consistent hyperparameters and data augmentation techniques to improve generalization.

Performance Evaluation

Model performance was evaluated quantitatively on the test set using standard metrics:

1. Precision – the accuracy of positive predictions
2. Recall – the ability to find all positive instances
3. Total Loss – a sum of classification, localization, and regularization losses during training

The best-performing model was then subjected to a further qualitative prediction stage involving 100 real-time samples to assess its practical utility and identify common failure modes.

The evaluation framework was designed not just to find the most accurate model, but to identify the most efficient model suitable for real-world deployment on common hardware like laptops or mid-range smartphones, aligning with the goal of creating accessible technology.

RESULT AND DISCUSSION

Comparative Model Performance

The quantitative results from the test set evaluation are summarized in Table I. The SSD MobileNet V2 model outperformed its counterparts across key metrics, achieving the highest precision (0.8869) and recall (0.8867). Notably, it also concluded with the lowest classification loss, indicating superior learning of gesture features.

TABLE I COMPARISON OF MODEL PERFORMANCE METRICS

Model	**SSD MobileNet V2**	**EfficientDet D0**	**SSD ResNet50**
Precision	0.8869	0.7085	0.805
Recall	0.8867	0.71	0.825
Classification loss	0.1179	0.1465	0.1737
Localization loss	0.02542	0.0029943	0.06509
Regularization loss	0.1168	0.02264	0.1166

The EfficientDet D0 model, while achieving the lowest localization loss, struggled with classification accuracy. The SSD ResNet50 model delivered middling performance. The superior results of the SSD MobileNet V2 model

can be attributed to its architecture, which is specifically designed for efficiency and speed without a significant sacrifice in accuracy on tasks well-suited to its feature extraction capabilities, making it ideal for this application.

Real-Time Prediction and Validation

The SSD MobileNet V2 model was deployed for real-time prediction on 100 new gesture samples. The system achieved an overall accuracy of 95.5%, with detailed performance per class shown in Table II and the confusion matrix in Fig. 4.

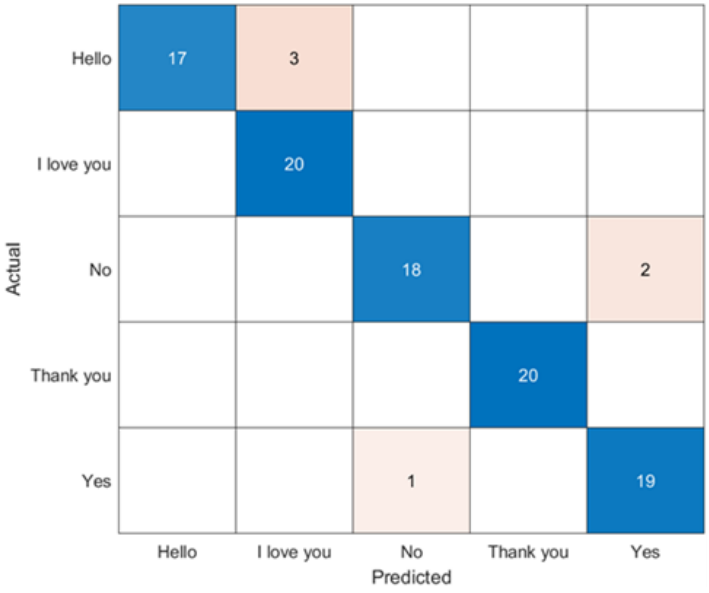

Fig. 4 Normalized Confusion Matrix for the real-time prediction stage on 100 samples

TABLE II Performance of Prediction

Gesture Class	Precision	Recall	F1-Score
Hello	100%	85%	91.9%
I love you	86.9%	100%	93.0%
Thank you	100%	100%	100%
Yes	90.5%	95%	92.7%
No	94.7%	90%	92.3%

The "Thank You" gesture was perfectly recognized. Most misclassifications occur between gestures with similar morphologies or motion paths. For instance:

- "Hello" (85% Recall): Was occasionally confused with "No," likely due to similar wrist orientation at certain motion phases.
- "Yes" (95% Recall) and "No" (90% Recall): Showed some mutual confusion, potentially due to the opposite nature of their movements leading to ambiguous mid-motion frames.

To further demonstrate the system's practical utility in a multi-user environment, a real-time prediction snapshot was captured (Fig. 5). The figure illustrates the system simultaneously processing gestures from three participants, showcasing its ability to handle multiple inputs in a realistic scenario. Notably, the model successfully recognized and distinguished between different gestures performed concurrently by the same individual, such as a "Hello" (84% confidence) and an "I Love You" (97% confidence) gesture by the first participant, and a "Yes" (95% confidence) and "No" (96% confidence) by the last participant. The prediction confidence for a "Thank You" gesture from a second user was 88%. This visual evidence strongly supports the quantitative findings, confirming that the model is not only accurate in controlled tests but also robust and effective in a dynamic, multi-person setting.

This capability for multi-user, multi-gesture recognition is a significant step towards a practical communication aid that could be used in group settings, moving beyond a simple one-user-one-system paradigm.

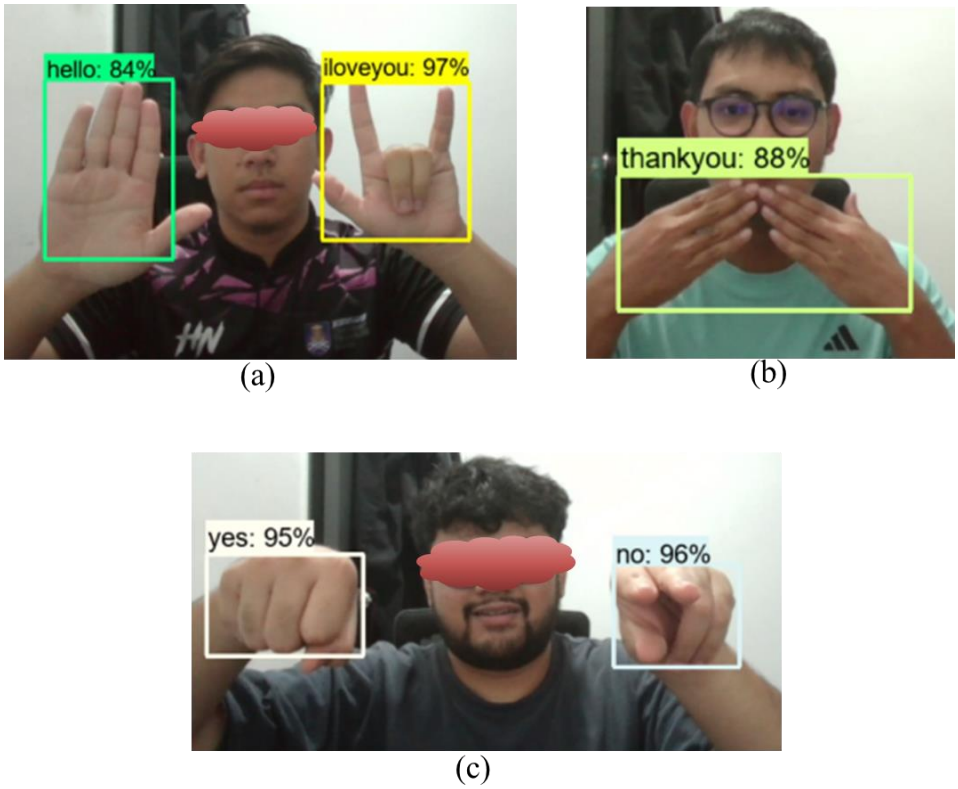

Fig. 4 Prediction stage showing the accuracy of individual test across different gesture classes

These findings are consistent with literature [13] and highlight a key challenge in gesture recognition: distinguishing between gestures with high visual similarity. Future work could incorporate temporal data to better disambiguate gestures based on their movement trajectory rather than a single frame.

A deeper analysis of the misclassifications reveals important insights into future development. The confusion between "Hello" and "No," for example, is not merely a technical error but points to a fundamental challenge in designing intuitive gestures for a diverse user base. A gesture that is clear and distinct for one user may be ambiguous from another angle or when performed with limited motor control, a common issue for individuals with certain neurological conditions. This highlights the necessity of user-centered design in assistive technology, where the target population must be involved in the selection and definition of gestures to ensure they are both distinguishable and physically achievable [14].

Furthermore, the high performance of the lightweight SSD MobileNet V2 model has significant sociotechnical implications. Its efficiency means it can be deployed on affordable, off-the-shelf smartphones, dramatically reducing the cost barrier associated with traditional high-end AAC devices. This democratization of technology can increase access to communication tools in low-resource settings and developing countries, aligning with global goals for inclusive development and disability rights [14],[15]. The model's ability to function in real-time is crucial for sustaining natural conversation rhythms.

CONCLUSION

This study successfully developed and evaluated a deep learning-based system for translating hand gestures into text, demonstrating a viable pathway for enhancing communication accessibility. The comparative analysis established that the SSD MobileNet V2 FPNLite 320x320 model provides the best balance of high accuracy (95.5% in real-time tests) and computational efficiency, making it ideally suited for deployment on low-cost, portable hardware.

The primary limitation of this work is its reliance on static image recognition, which can struggle with dynamic gestures and subtle variations in motion. Furthermore, the dataset, while diverse in angles, was limited in size and demographic variability. The use of a webcam for data collection, while accessible, also limits the variability in image quality and environmental conditions compared to a multi-camera or professionally curated dataset.

Future research will focus on several key areas:

1. Increasing the number of images and including participants from diverse backgrounds to improve model robustness and equity.
2. Implementing video-based models to capture the sequential nature of gestures, thereby improving accuracy for motion-dependent signs.
3. Integrating the optimal model into a standalone application or embedded system for field testing with target users, which is the ultimate measure of its success as an assistive technology. This includes adding a text-to-speech component to read the translated text aloud, further enhancing accessibility.
4. Techniques like quantization and pruning could be applied to the chosen model to further reduce its size and latency, enabling seamless operation on even less powerful devices.
5. A crucial next step involves formal collaboration with speech-language pathologists and individuals with speech impairments to co-design the gesture set and user interface, ensuring the technology is not only effective but also empowering and tailored to real-world needs.

In conclusion, this work affirms the significant potential of lightweight deep learning models in creating practical and effective assistive communication tools, offering a tangible step toward greater social inclusion and independence for individuals with speech disabilities.

ACKNOWLEDGMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) and Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer (FTKEK) for the support.

REFERENCES

1. Blasko, G., 2025. Unveiling underlying systemic isolation challenges for AAC users. Augmentative and Alternative Communication, 41(3), pp.215-222.
2. Hasegawa-Johnson, M., Zheng, X., Kim, H., Mendes, C., Dickinson, M., Hege, E., Zwilling, C., Channell, M.M., Mattie, L., Hodges, H. and Ramig, L., 2024. Community-supported shared infrastructure in support of speech accessibility. Journal of Speech, Language, and Hearing Research, 67(11), pp.4162-4175.
3. Al-Qurishi, M., Khalid, T. and Souissi, R., 2021. Deep learning for sign language recognition: Current techniques, benchmarks, and open issues. IEEE Access, 9, pp.126917-126951.
4. Mohamed, A.S., Hassan, N.F. and Jamil, A.S., 2024. Real-Time Hand Gesture Recognition: A Comprehensive Review of Techniques, Applications, and Challenges. Cybernetics and Information Technologies, 24(3), pp.163-181.
5. Cui, C., Sunar, M.S. and Su, G.E., 2025. Deep vision-based real-time hand gesture recognition: a review. PeerJ Computer Science, 11, p.e2921.
6. [Griffiths, T., Slaughter, R. and Waller, A., 2024. Use of artificial intelligence (AI) in augmentative and alternative communication (AAC): community consultation on risks, benefits and the need for a code of practice. Journal of Enabling Technologies, 18(4), pp.232-247.
7. Al Farid, F., Hashim, N., Abdullah, J., Bhuiyan, M.R., Shahida Mohd Isa, W.N., Uddin, J., Haque, M.A. and Husen, M.N., 2022. A structured and methodological review on vision-based hand gesture recognition system. Journal of Imaging, 8(6), p.153.
8. Arafah, M., Achmad, A. and Areni, I.S., 2019, October. Face recognition system using Viola Jones, histograms of oriented gradients and multi-class support vector machine. In Journal of Physics: Conference Series (Vol. 1341, No. 4, p. 042005). IOP Publishing.
9. Kumar, A. and Srivastava, S., 2020. Object detection system based on convolution neural networks using single shot multi-box detector. Procedia Computer Science, 171, pp.2610-2617.

10. Varnima, E.K. and Ramachandran, C., 2020, June. Real-time Gender Identification from Face Images using you only look once (yolo). In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184) (pp. 1074-1077). IEEE.
11. Dong, K., Zhou, C., Ruan, Y. and Li, Y., 2020, December. MobileNetV2 model for image classification. In 2020 2nd International Conference on Information Technology and Computer Application (ITCA) (pp. 476-480). IEEE.
12. Han, X., Lu, F. and Tian, G., 2022. Efficient 3D CNNs with knowledge transfer for sign language recognition. Multimedia Tools and Applications, 81(7), pp.10071-10090.
13. Jagtap, S., Jadhav, K., Temkar, R. and Deshmukh, M., 2024. Real-time Sign Language Recognition Using MobileNetV2 and Transfer Learning. arXiv preprint arXiv:2412.07486.
14. Blasko, G., Light, J., McNaughton, D., Williams, B. and Zimmerman, J., 2025. Nothing about AAC users without AAC users: A call for meaningful inclusion in research, technology development, and professional training. Augmentative and Alternative Communication, 41(3), pp.184-194.
15. World Health Organization, 2024. Health equity for persons with disabilities: guide for action. World Health Organization.