Advancing Sign Language Recognition: A YOLO v.11-Based Deep Learning Framework for Alphabet and Transactional Hand Gesture Detection

Abdelrahman T. Elgohr [1], Mohamed S. Elhadidy [1], Marwa El-geneedy [1], Shimaa Akram [2], and Mahmoud A. A. Mousa [3]

[1] Department of Mechatronics Engineering, Faculty of Engineering, Horus University, New Damietta 34517, Egypt
[2] Communications and Electronics Engineering Dept., Faculty of Engineering, Horus University Egypt, New Damietta, Egypt
[3] School of Mathematical and Computer Sciences, Heriot Watt University, Dubai, UAE
atarek@horus.edu.eg; melhadidy@horus.edu.eg; melgenedy@horus.edu.eg; ssoliman@horus.edu.eg; m.mousa@hw.ac.uk

Abstract

Sign language recognition is an essential tool that facilitates communication for those with hearing and speech disabilities. Conventional recognition techniques frequently encounter challenges in re-al-time performance, resilience, and accuracy owing to fluctuations in hand positions, backdrops, and lighting conditions. This paper presents a YOLO v11-based deep learning system for recognizing ASL, concentrating on both alphabetic and transactional hand motions to mitigate existing constraints. The model is engineered to function in real-time while ensuring high precision and resilience across varied contexts. The methodology adheres to a systematic pipeline, commencing with dataset gathering and preprocessing, which include image augmentation, normalization, and scaling to guarantee model generalization. The YOLOv11 architecture utilis-es an improved backbone, neck, and detecting head for effective feature extraction and classification. Training is enhanced by the utilisation of the AdamW optimizer, a meticulously adjusted learning rate, and a loss function that integrates box loss, classification loss, and distribution focal loss. Performance is assessed using precision, recall, mAP, and inference rate to guarantee the model's accuracy and efficiency. Experimental findings indicate that the suggested model attains 95.4% precision, 94.8% recall, and 98.1% mAP, markedly surpassing conventional methods. The amalgama-tion of GRAD-CAM with occlusion sensitivity significantly im-proves model interpretability. This research offers a robust and scalable approach for real-time sign language detection, facilitating enhanced accessibility in communication technologies, and interactive systems.

Introduction

Those with hearing and speech difficulties use sign languages to express thoughts, feelings, and information through manual gestures, or facial expressions. Sign language unites deaf and hearing people worldwide, encouraging inclusivity and understanding in social, educational, and professional settings (Najib, 2024; Wadhawan & Ku-mar, 2020). Sign languages have developed separately in different countries, resulting in unique languages like ASL, BSL, ISL, and ArSL. Despite their differences, all sign languages improve communication for hearing and speech-impaired people (Alsharif et al., 2023; Elgohr et al., 2025).

Communication accessibility requires sign language motion classification. Precision and effectiveness in sign language identification can help create applications in education, healthcare, customer service, and public services. These advancements illuminate gesture-based communication, which may improve human-computer interaction, robotics, and augmented reality. Resilient classification techniques enable real-time translation systems for sign language users and non-users without interpreters (Noor et al., 2024; Y. Zhang & Jiang, 2024).

Classification categorizes data by traits and patterns. Sign language recognition uses classification to match hand forms, movements, and gestures to letters, words, and sentences. Figure 1 shows that classification is most commonly used in real-time translation, sign language education, hearing aids, and customer service and healthcare communication interfaces (Mousa et al., 2024).

Recent deep learning advances have changed sign language dataset classification. Advanced models that can detect subtle spatial and temporal patterns in gesture images and videos have replaced manual feature design and statistical frameworks in machine learning. CNNs efficiently extract spatial features from images, while LSTM networks efficiently capture temporal relationships in sequential data. Transformer-based architectures using self-attention approaches are becoming popular for handling long-range dependencies (Baihan et al., 2024; Haque et al., 2023).

YOLO (You Only Look Once) models are popular in object identification, particularly sign language recognition. YOLO v.11 improves processing speed and accuracy, mak-

ing it a viable real-time sign language identification solution. YOLO models partition images into grids and forecast bounding boxes and class probabilities using a single neural network. This improved technique reduces computational load, making YOLO models suitable for edge devices and mobile apps. Hybrid CNN-attention mechanism-recurrent layer models can capture static and dynamic hand motions (Ali et al., 2025; Elgohr et al., 2024).

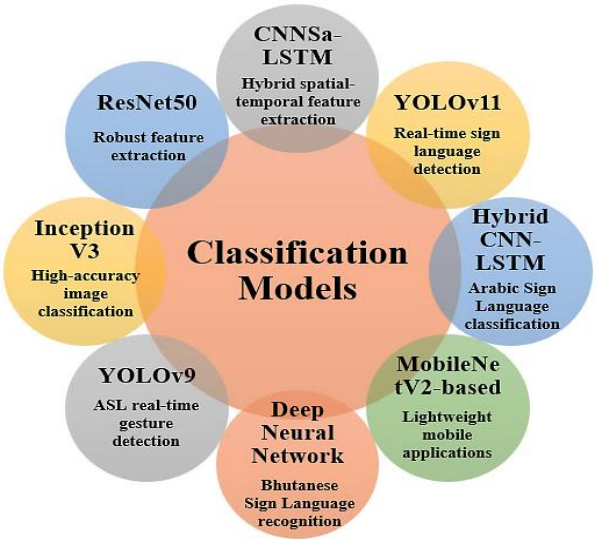

Figure 1: Most common usage of recent classification models.

Related Works

Sign language recognition has experienced considerable progress over the years, mostly due to the incorporation of deep learning methodologies. This sub-section examines the current literature, emphasizing significant contributions, techniques, and issues encountered in diverse investigations. Sign language recognition systems can be classified into vision-based and sensor-based approaches. Vision-based systems, preferred for their simplicity and cost-efficiency, employ computer vision algorithms to interpret hand movements, whereas sensor-based methods depend on wearable equipment such as gloves integrated with sensors.

Aksoy et al., 2021, investigated the identification of Turkish Sign Language (TSL) via CNN-based architectures. Their research utilised a bespoke dataset including 10,223 photos depicting 29 letters. The researchers utilised multiple models, such as CapsNet, AlexNet, ResNet-50, and TSLNet. CapsNet and TSLNet attained peak accuracies of 99.7% and 99.6%, respectively. The research illustrated the efficacy of data augmentation and picture preprocessing methods such as filtering and segmentation to enhance recognition accuracy. Haque et al., (2023) utilised CNNs for Bangladeshi Sign Language (BdSL) identification, with 99% accuracy in training and 93% in testing. Both studies highlighted dataset customization and preprocessing as essential elements for performance improvement.

Buttar et al., 2023, and Baihan et al., 2024, utilised hybrid deep learning methodologies that integrate CNNs with LSTM layers. Buttar et al. created a model employing LSTM and YOLOv6 for American Sign Language (ASL) recognition, attaining 96% accuracy for static signs and 92% for dynamic signs. Baihan et al. presented CNNSa-LSTM via a hybrid optimisation method, attaining an accuracy of 98.7%. The primary difference between these models is Baihan et al.'s implementation of an innovative hybrid optimizer that integrates HOA and PFA for feature extraction optimisation.

Wadhawan and Kumar, 2020, and J. Zhang et al., 2024, investigated static sign recognition employing CNNs. Wadhawan's CNN model for Indian Sign Language (ISL) attained 99.90% accuracy using greyscale photos, whilst Zhang's Dual-Path Background Erasure Convolutional Neural Network (DPCNN) achieved 99.52% accuracy by eliminating background characteristics to improve recognition precision. Both methodologies underscored the importance of feature extraction techniques; however, Zhang's approach presented a distinctive dual-path architecture for enhanced background management.

Kyaw et al., 2024, concentrated on the recognition of Myanmar Sign Language (MSL) employing CNN models. They attained 99% accuracy by integrating movies recorded under different lighting conditions, highlighting the need of varied training datasets for enhanced model generalization. Alyami et al., 2024, utilised transformer-based models for Arabic Sign Language (ArSL), attaining an accuracy of 99.74% and highlighting the significance of non-manual face aspects in gesture detection.

(Likhar et al., 2020, attained elevated recognition accuracy for Indian Sign Language (ISL) by the utilisation of RGB-D data with convolutional neural networks (CNNs), achieving 98.81% and 99.08% accuracy for static and dynamic motions, respectively. Das et al., 2023, introduced a CNN-BiLSTM model for word-level recognition using keyframe extraction, attaining an accuracy of 87.67%. The comparison reveals that Likhar's model outperformed in both static and dynamic recognition tests, whereas Das's model prioritized efficiency in processing time-series data.

Vetagiri et al., 2025, examined CNN-BiLSTM hybrids for sequential classification tasks, offering insights pertinent to sign language recognition. While their research primarily focused on detecting sexism, the approaches described could be used to gesture-based classification problems because of their efficacy in capturing temporal relationships.

The analyzed studies emphasize the efficacy of CNNs, LSTMs, and hybrid models in sign language recognition. Models utilising spatial and temporal characteristics, such as CNNSa-LSTM and hybrid YOLOv6-LSTM methodologies, have demonstrated encouraging outcomes across multiple languages. The implementation of background erasure

techniques and transformer-based models significantly improves performance, particularly in real-time applications. This study seeks to enhance existing methodologies by classifying a bespoke dataset for sign language recognition utilising YOLO v.11, integrating both static and dynamic gestures to augment classification accuracy and broaden the applicability of sign language recognition in more practical and interactive contexts.

Table 1 provides a clear comparison of the techniques, languages, and accuracies achieved by each study, offering a concise overview of the field's advancements and areas of focus.

Material & Methods

Figure 2 delineates the approach of a hand gesture recognition system utilising YOLOv11, detailing the sequential procedure from data collection to final prediction. The method commences with the input stage, during which hand gesture images are gathered. During the data pre-processing phase, these images are subjected to labelling, partitioning, and augmentation to enhance model generalization. Upon preparation, the data is transmitted through the YOLOv11 layers, which include a input layer for feature extraction, a backbone layer for advanced feature analysis, a neck layer for enhancing feature maps, and a detection head for predicting bounding boxes and class labels. The ReLU activation function is utilised to introduce non-linearity, hence improving learning efficiency.

Subsequently, the model training phase enhances detection accuracy through the integration of a loss function, an optimizer, modifications to the learning rate, and mAP (mean average precision) for assessment. To enhance model transparency, explainability techniques like GRAD-CAM and occlusion sensitivity are employed, facilitating the visualization of significant elements that affect predictions. The output stage provides the final prediction, including bounding boxes and class labels, to ensure precise hand gesture identification. This systematic approach combines sophisticated deep learning methods with explanatory aids, rendering the system both resilient and comprehensible.

The methodology section details the dataset acquisition, preprocessing techniques, and the architecture of the YOLOv11 model, including the training process and explainability techniques. The results and discussion section presents the experimental outcomes, comparing the model's performance with state-of-the-art techniques and analyzing its superiority. Subsequently, the conclusion and future work section summarizes the key contributions of the study while outlining potential enhancements such as multi-language recognition, mobile deployment, and integration with other modalities. The paper concludes with references, ensuring a well-documented and academically rigorous study.

Ref.	Language	Model	Key Techniques	Accuracy	Dataset Size	Hardware
(Aksoy et al., 2021)	Turkish Sign Language	CapsNet/TSLNet	Data Augmentation, CNN	99.7% - 99.6%	10,223 images	GPU
(Buttar et al., 2023)	American Sign Language	LSTM+YOLOv6	Skeleton-based Features	96% - 92%	Custom dataset	GPU
(J. Zhang et al., 2024)		DPCNN	Background Erasure	99.52%	ASL Finger Spelling	CPU
(Wadhawan & Kumar, 2020)	Indian Sign Language	CNN	50 CNN Models Tested	99.90%	35,000 images	GPU
(Likhar et al., 2020)		CNN	RGB-D Data	98.81% - 99.08%	RGB-D dataset	GPU
(Das et al., 2023)		CNN-BiLSTM	Key-frame Extraction	87.67%	Custom dataset	GPU
(Haque et al., 2023)	Bangladeshi Sign Language	CNN	DenseNet201+ResNet50-V2	99% - 93%	992 images	CPU
(Kyaw et al., 2024)	Myanmar Sign Language	CNN	Diverse Lighting Conditions	99%	Video dataset	GPU
(Alyami et al., 2024)	Arabic Sign Language	Transformer-based	MediaPipe Keypoints	99.74%	KArSL-100	GPU
(Baihan et al., 2024)	Multiple Languages	CNNSa-LSTM	HOA+PFA Optimizer	98.7%	Multiple datasets	GPU
(Vetagiri et al., 2025)	Text-based Classification Task	CNN-BiLSTM	Temporal Dependency Analysis	92%	Multi Hate	GPU

Table 1: Comparative Summary of Related Research.

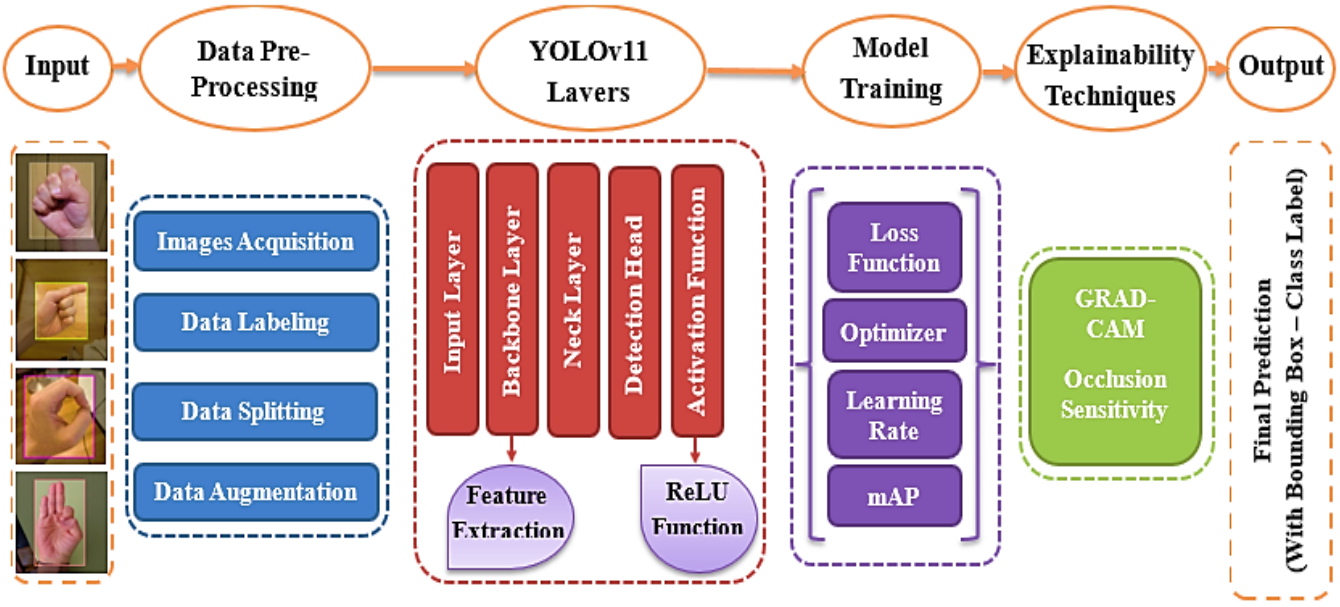

Figure 2. Proposed system architecture.

Dataset Description

This study used The American Sign Language (ASL) dataset which was sourced from Roboflow Universe/duyguj/American-sign-language-letters. All images, as in Figure 3, in the dataset were pre-labeled, ensuring accurate training data. Additionally, data augmentation techniques were applied within Roboflow to increase the variability of the dataset, improving the model's generalization. Techniques such as flipping, rotation, and brightness adjustments were employed. This dataset contains a total of 1224 images, which are split into three sets 82% (1008 images) for Train set, 12% (144 images) for Validation set, and 6% (72 images) for Test set.

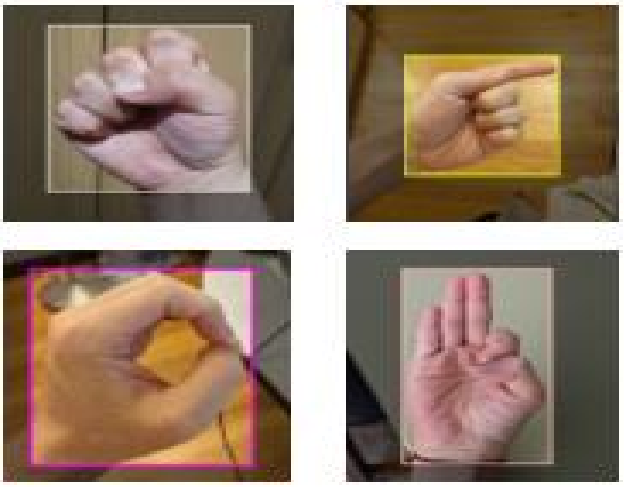

Figure 3. Dataset samples

Dataset Pre-Processing

Preprocessing plays a crucial role in enhancing the accuracy and reliability of the YOLO11 model for real-time sign language detection. This study incorporated various preprocessing techniques to ensure that the inputs for training and inference are of high quality. First, we applied Auto-Orientation to ensure that all images were properly aligned, preventing any misinterpretation caused by rotation inconsistencies.

All input images were resized to 640×640 pixels, the requisite input dimension for YOLO11, to ensure consistency and enhance performance. To improve model generalization, several data augmentation techniques were utilized on the dataset. The images underwent rotations ranging from -15° to +15° to represent different hand positions, and brightness levels were modified between -10% and +10% to reflect various lighting scenarios. Furthermore, a 2px Gaussian blur was applied to create motion blur and replicate real-world camera irregularities. To improve model stability during training, we normalize pixel values to a range between 0 and 1. For video preprocessing, we extracted frames in real-time to facilitate sequential gesture recognition.

We also employed edge detection and background subtraction methods to enhance the clarity of hand gestures. Finally, we integrated OpenCV to preprocess the video frames before sending them to the YOLO11 model. These preprocessing techniques guaranteed that the YOLO11 model was provided with high-quality, well-augmented, and standardized inputs, enhancing its ability to recognize sign language gestures across different conditions.

YOLO v.11 Model

YOLO v.11 enhances the developments of its predecessors (YOLO9, YOLO10) by including an upgraded architecture, improved feature extraction, and refined training methodologies (Huang et al., 2024). The model is capable of managing various activities across diverse domains and has strong scalability, accommodating both mobile CPUs and robust GPUs (Bakirci et al., 2024; El-geneedy et al., 2024). YOLO11 can be obtained in five different sizes, with parameter counts varying from 2.6 million to 56.9 million, and attains MAP scores ranging from 39.5 to 54.7 on the COCO dataset, used for initial pre-training (Alkhammash, 2025; Hassan et al., 2024). This study used the compact version of YOLO11 for the efficient real-time detection of hand signs. In addition to its powerful real-time object identification capabilities, this model performs in a variety of other tasks, including classification, instance segmentation, pose estimation, semantic segmentation, and object detection. Recognizing hand gestures was a breeze using the compact version of YOLO11 for object identification in this work. It effortlessly handled real-time sign recognition.

It was required to refine the YOLOv11 model using the ASL dataset in order for it to learn object detection tailored specifically for sign language. The training process applied dataset augmentation via Roboflow, which helped the model become more resilient by improving the dataset with different transformations. The YOLOv11 model was first trained on the augmented dataset and then validated with a separate set of data to monitor the learning process and avoid overfitting. After training, the model was tested on a dedicated test set which showcased its competency in predicting unseen data, successfully aiding in recognizing sign language gestures.

The YOLOv11 model architecture is designed to effectively identify American Sign Language letters with a tripartite structure consisting of the backbone, neck, and head. The backbone acts as the feature extractor, using convolutional layers with residual connections to collect low-level and high-level image characteristics. It systematically reduces the image's dimensions while augmenting the depth of feature maps to retrieve intricate details. The neck component links the backbone to the detecting head, using SPPF (Spatial Pyramid Pooling Fast) layers to consolidate multi-scale data and improve spatial information. Furthermore, the C2PSA (Cross-Stage Partial Spatial Attention) blocks enhance attention processes, enabling the model to concentrate on pertinent areas of hand motions. The detection head has three output layers that forecast item bounding boxes, class probabilities, and confidence ratings at various scales. This multi-scale detection method enhances the model's capacity to identify gestures of diverse sizes and orientations. The design facilitates Automatic Mixed Precision (AMP), enhancing training speed without sacrificing precision. Residual connections and attention mechanisms assist in reducing information loss during feature extraction. YOLOv11's design effectively balances speed, accuracy, and computing economy, making it highly suitable for sign language identification applications.

Our training procedure was set up with comparable hyperparameters: an optimizer called AdamW with the following settings: a learning rate of 0.000333, momentum=0.9, parameter groups 81 weight (decay=0.0), 88 weight (decay=0.0005), 87 bias (decay=0.0), 16 batches, and 10 epochs in total. All training was performed only on Kaggle, implementing its high-performance computing infrastructure to accommodate the time-consuming nature of these challenges. A warmup phase of 3 epochs is applied with lower learning rates to prevent sudden weight updates and stabilize training. The overlapping mask prediction technique is applied with a mask ratio of 4 to improve segmentation quality. The dropout rate is set to 0.0 to maintain consistent feature extraction. The model uses Rectified Linear Units (ReLU) as activation functions in convolutional layers. Batch normalization is applied after each convolutional layer to stabilize learning. This combination of layers and parameters allows the model to achieve high detection accuracy with efficient computation. The loss function combines box loss, classification loss, and distribution focal loss (DFL) to ensure accurate localization and classification of hand gestures. Table 2 presents the model layers and parameters it consists of 319 layers with 2,594,910 parameters and 6.5 GFLOPs (Giga Floating Point Operations).

Layer	Parameters	Description
Conv Layer 1	464	3x3 Convolution with 16 filters (stride 2)
Conv Layer 2	4672	3x3 Convolution with 32 filters (stride 2)
C3k2 Block 1	6640	Cross-Stage Partial Module with 64 filters
SPPF Layer	164608	Spatial Pyramid Pooling layer
C2PSA	249728	Cross-Stage Partial with Spatial Attention
Detection Head	435742	Final Detection layers with 3 scale outputs

Table 2: Model layers and parameters.

For explainability, the model generates Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations to highlight the most influential regions in an image for each prediction. The Grad-CAM maps help interpret the model's decision-making process by showing which parts of the hand gesture contribute most to the classification result. Additionally, bounding boxes with confidence scores are plotted on the images to visualize the model's predictions. The model

uses Automatic Mixed Precision (AMP) to speed up training while maintaining numerical stability. These training and explainability techniques provide insights into the model's decision-making process and improve its performance for sign language detection tasks.

Performance Metrics

In the improved American Sign Language (ASL) recognition model, the performance of the system is rated with standard measures commonly applied in object detection and classification problem. How accurate the model is to recognize and classify hand signs is measured by a combination of Intersection over Union (IoU), true positive (TP), false positive (FP), false negative (FN), precision, recall, mean Average Precision (mAP), and inference rate (Elhadidy et al., 2025; Maurício et al., 2023).

The value of IoU is significant in measuring the overlap between the predicted bounding box and the ground truth bounding box. For this study, a threshold of 0.5 is employed for IoU, meaning that a prediction is considered correct if the overlap between the two boxes is greater than 50%. In this way, good localizations of the hand signs are only considered successful detection (Dai & Fang, 2025).

True positive (TP) occurs when the predicted bounding box correctly identifies the hand sign within the true bounding box. False positive (FP) occurs when the model makes a prediction in a bounding box where there is no hand sign, which gives rise to false detections. False negative (FN) occurs when the model fails to make a prediction despite having a hand sign within the true bounding box (Macsik et al., 2024). In this study, true negative (TN) is not considered since the dataset lacks any images except for hand signs.

Dataset split, which is 82% training, 12% validation, and 6% testing, gives a good evaluation of the model's ability to generalize to new data. We use the Keras and Torch framework to implement it. It is necessary to use parallel processing while training deep neural networks. Since this was the case, we used the open-source software Python 3.0 and Kaggle to train and evaluate the classifiers. Additionally, we utilized NVIDIA TESLA P100 graphics processing units (GPUs) and 16 GB of RAM.

Precision

This metric focuses on true positives and false positives. High precision is achieved when false positives are low. It is calculated as (Juba & Le, 2019):

$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

Sensitivity (Recall)

This metric highlights true positive and false negatives. High sensitivity is achieved when false negatives are low. It is calculated as (Erickson & Kitamura, 2021):

$$Sensitivity = \frac{Truepositives}{True\ positives + False\ negatives}$$

Mean Average Precision (mAP)

Mean Average Precision (mAP) is another popular object detection measure that calculates the average precision over recall levels from 0 to 1. It provides a single numeric value representing the overall detection accuracy of the model for all hand signs (Juba & Le, 2019):

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

Where, N is the number of classes and AP_i is the average precision for each class (Usama et al., 2021).

Inference rate, in frames per second (fps) or milliseconds (ms) per frame, is another essential metric for real-time systems. Higher fps indicates that the model can process more images per second, which is very important for real-time sign language recognition (Elgohr, Mousa, et al., 2025).

Through these performance metrics, the accuracy, efficiency, and responsiveness of the ASL detection model are thoroughly tested to validate its reliability in real-time applications.

Result & Discussion

The results of the YOLO11 model trained for American Sign Language (ASL) letter detection can be said to have been promising over the span of 100 epochs within the set limits. In Figure 4. The probability of correct detection is shown in the bounding box. The model's performance was evaluated with distinct parameters as shown in Figure 5. The YOLO11 model achieved 95.4% for the precision, 94.8% for the recall, and 98.1% for mAP.

Also, Throughout the training, the loss metrics all improved steadily over time. The box loss decreases from approximately 0.35 to nearly 0.12, reflecting better accuracy in the bounding box predictions of the model. The cls loss decreases from 1.6 to nearly 0.4, reflecting significant improvement in the model's ability to properly classify objects in the predicted boxes. The dfl loss also decreases from 1.05 to approximately 0.9, reflecting better localization performance. Next Throughout the validation, box loss starts at around 1.5 and decreases to nearly 0.2, cls loss falls drastically from around 5 to nearly 1, whereas dfl loss falls from 2.5 to around 0.5, showing how well the model can generalize on new unseen data. Even though the overall performance is promising, some misclassifications to detect were evident particularly for characters whose hand shapes are the same as shown in the confusion matrix in Figure 6.

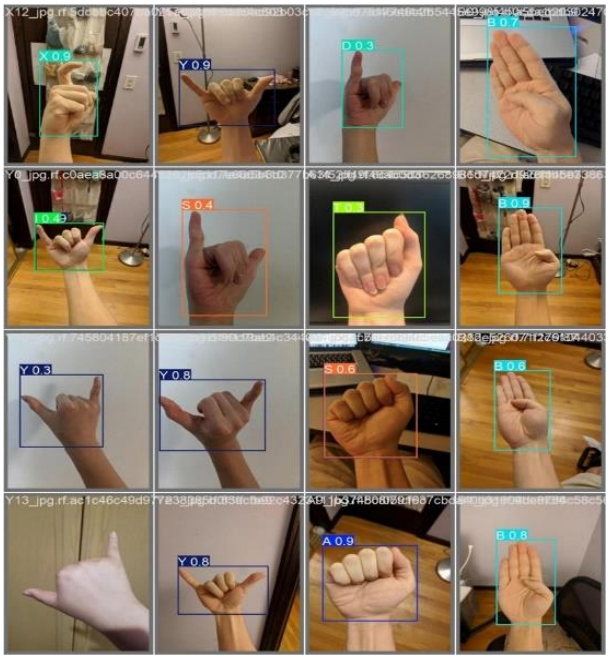

Figure 4: The predictions of the model and its confidence score on different signs.

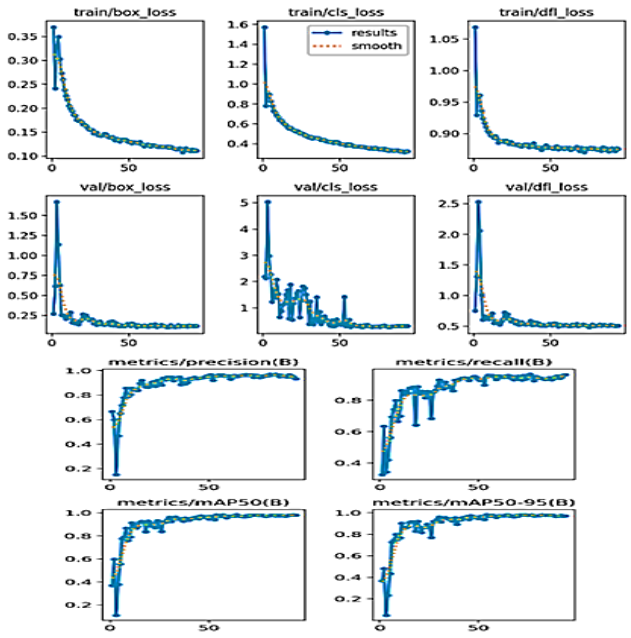

Figure 5. The training validation precision, recall, mAP50, mAP50-90 and different types of loss of the Proposed model

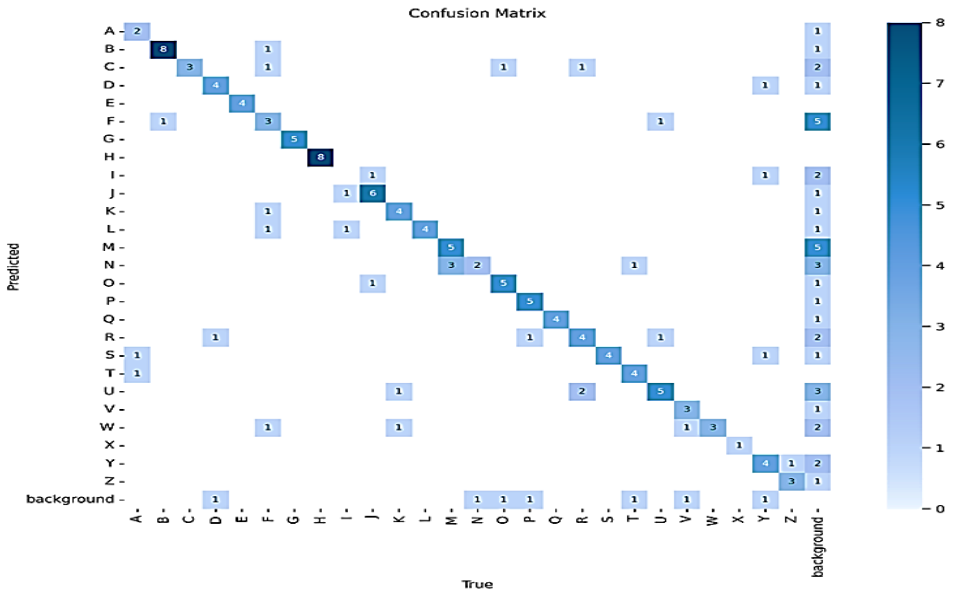

Figure 6. Confusion matrix of the model results.

The experimental results demonstrate the superiority of the proposed YOLOv11-based sign language recognition system compared to traditional approaches. The model achieved 95.4% precision, 94.8% recall, and 98.1% mAP, highlighting its high accuracy in detecting and classifying both alphabetical and transactional hand gestures. These results outperform conventional CNN-based models and previous YOLO versions, which often suffer from misclassifications due to variations in background, lighting, and hand orientations. The optimized YOLOv11 architecture, featuring an enhanced backbone, neck, and detection head, enables precise feature extraction and classification, ensuring reliable predictions in real-time scenarios. Furthermore, the incorporation of explainability techniques such as GRAD-CAM and occlusion sensitivity allows for deeper insights

into the model's decision-making process, reinforcing its trustworthiness and interpretability in real-world applications.

The proposed system offers several advantages over existing sign language recognition models. Firstly, it ensures real-time performance with minimal computational overhead, making it suitable for embedded systems, mobile devices, and edge computing applications. Secondly, the data augmentation and preprocessing techniques, including normalization, resizing, and background filtering, enhance model robustness, ensuring consistent performance across different lighting conditions and hand positions. Additionally, the integration of an adaptive loss function and an optimized training strategy using the AdamW optimizer improves convergence speed and classification accuracy. The system's flexibility allows for scalability, making it applicable to a wide range of sign languages beyond ASL. With these strengths, the proposed model has the potential to enhance communication accessibility, benefiting individuals with hearing impairments and paving the way for further advancements in assistive technologies and human-computer interaction.

Conclusion & Future Work

This study introduced a YOLOv11-based deep learning system for real-time recognition of American Sign Language (ASL), emphasizing both alphabetic and transactional hand motions. The system exhibited exceptional accuracy and robustness across many environmental circumstances by utilising a meticulously selected dataset, sophisticated data augmentation methods, and an optimised YOLOv11 architecture. The model attained 95.4% precision, 94.8% recall, and 98.1% mean Average Precision (mAP), exceeding conventional techniques in sign language identification. The incorporation of GRAD-CAM and occlusion sensitivity elucidated the model's decision-making process, hence improving interpretability and dependability for practical applications. The results underscore the effectiveness and scalability of the suggested system, rendering it a feasible solution for assistive technologies, educational instruments, and human-computer interaction interfaces.

Notwithstanding its commendable performance, there remain aspects for enhancement that subsequent efforts should tackle. Initially, augmenting the dataset to encompass a broader range of sign languages and dynamic gestures would enhance generalization across various linguistic situations. Furthermore, integrating transformer-based models with YOLOv11 may improve feature extraction for intricate gesture sequences. Subsequent research should concentrate on creating lightweight iterations of the model tailored for implementation on edge devices and mobile platforms, hence enhancing the accessibility of sign language recognition. Furthermore, the use of multi-modal data, including facial expressions and hand gestures, may enhance recognition precision in intricate real-world situations. Ultimately, real-time implementation and user evaluations must be performed to evaluate the model's usability and efficacy in interactive communication systems.

By tackling these challenges, forthcoming progress in deep learning and computer vision will further improve the accessibility and efficacy of sign language recognition systems, closing the communication divide between hearing and non-hearing individuals across diverse societal and technological spheres.

References

Aksoy, B., Salman, O. K. M., & Ekrem, Ö. 2021. Detection of Turkish Sign Language Using Deep Learning and Image Processing Methods. Applied Artificial Intelligence, 35(12), 952–981. https://doi.org/10.1080/08839514.2021.1982184

Ali, A. O., Elgohr, A. T., El-Mahdy, M. H., Zohir, H. M., Emam, A. Z., Mostafa, M. G., Al-Razgan, M., Kasem, H. M., & Elhadidy, M. S. 2025. Advancements in photovoltaic technology: A comprehensive review of recent advances and future prospects. Energy Conversion and Management: X, 26. https://doi.org/10.1016/j.ecmx.2025.100952

Alkhammash, E. H. 2025. Multi-Classification Using YOLOv11 and Hybrid YOLO11n-MobileNet Models: A Fire Classes Case Study. Fire, 8(1), 17. https://doi.org/10.3390/fire8010017

Alsharif, B., Altaher, A. S., Altaher, A., Ilyas, M., & Alalwany, E. 2023. Deep Learning Technology to Recognize American Sign Language Alphabet. Sensors, 23(18). https://doi.org/10.3390/s23187970

Alyami, S., Luqman, H., & Hammoudeh, M. 2024. Isolated Arabic Sign Language Recognition Using a Transformer-based Model and Landmark Keypoints. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1–19. https://doi.org/10.1145/3584984

Baihan, A., Alutaibi, A. I., Alshehri, M., & Sharma, S. K. 2024. Sign language recognition using modified deep learning network and hybrid optimization: a hybrid optimizer (HO) based optimized CNNSa-LSTM approach. Scientific Reports, 14(1).https://doi.org/10.1038/s41598-024-76174-7

Bakirci, M., Dmytrovych, P., Bayraktar, I., & Anatoliyovych, O. 2024. Multi-Class Vehicle Detection and Classification with YOLO11 on UAV-Captured Aerial Imagery. 2024 IEEE 7th International Conference on Actual Problems of Unmanned Aerial Vehicles Development (APUAVD), 191–196. https://doi.org/10.1109/APUAVD64488.2024.10765862

Buttar, A. M., Ahmad, U., Gumaei, A. H., Assiri, A., Akbar, M. A., & Alkhamees, B. F. 2023. Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs. Mathematics, 11(17). https://doi.org/10.3390/math11173729

Dai, Y., & Fang, X. 2025. An Armature Defect Self-Adaptation Quantitative Assessment System Based on Improved YOLO11 and the Segment Anything Model. Processes, 13(2), 532. https://doi.org/10.3390/pr13020532
Das, S., Biswas, S. Kr., & Purkayastha, B. 2023. A deep sign language recognition system for Indian sign language. Neural Computing and Applications, 35(2), 1469–1481. https://doi.org/10.1007/s00521-022-07840-y
El-geneedy, M., Elgohr, A. T., Elhadidy, M. S., & Akram, S. 2024. Early Lung Cancer Detection with a Fusion of Inception V3and Vision Transformers: A Binary Classification Study. 2024 International Conference on Future Telecommunications and Artificial Intelligence (IC-FTAI), 1–6. https://doi.org/10.1109/IC-FTAI62324.2024.10950031
Elgohr, A. T., Elhadidy, M. S., Elazab, M., Ahmed Hegazii, R., & El Sherbiny, M. M. 2024. Multi-Classification Model for Brain Tumor Early Prediction Based on Deep Learning Techniques. Journal of Engineering Research, 8, 2024. https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1646&context=erjeng
Elgohr, A. T., Khater, H. A., & Mousa, M. A. A. 2025. Trajectory optimization for 6 DOF robotic arm using WOA, GA, and novel WGA techniques. Results in Engineering, 25. https://doi.org/10.1016/j.rineng.2025.104511
Elgohr, A. T., Mousa, M. A. A., Mohamed, A. R., Khater, H. A., Ma'arif, A., & Suwarno, I. (2025). The Intelligence Behind Robotic Arms: A Deep Dive into Control Evolution. Journal of Robotics and Control (JRC), 6(3), 1478–1501. https://doi.org/10.18196/jrc.v6i3.25604
Elhadidy, M. S., Elgohr, A. T., El-geneedy, M., Akram, S., & Kasem, H. M. 2025. Comparative analysis for accurate multi-classification of brain tumor based on significant deep learning models. Computers in Biology and Medicine, 188. https://doi.org/10.1016/j.compbiomed.2025.109872
Erickson, B. J., & Kitamura, F. 2021. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiology: Artificial Intelligence, 3(3), e200126. https://doi.org/10.1148/ryai.2021200126
Haque, A., Pulok, R. A., Rahman, M. M., Akter, S., Khan, N., & Haque, S. 2023. Recognition of Bangladeshi Sign Language (BdSL) Words using Deep Convolutional Neural Networks (DCNNs). Emerging Science Journal, 7(6), 2183–2201. https://doi.org/10.28991/ESJ-2023-07-06-019
Hassan, O. H., Elhadidy, M. S., Elgohr, A. T., Ali, A. O., El-Mahdy, M. H., & Zaki, M. 2024. Performance Evaluation of a Standalone Solar-Powered Irrigation System in Desert Regions Using PVsyst: A Comprehensive Analysis. 2024 25th International Middle East Power System Conference (MEPCON), 1–6. https://doi.org/10.1109/MEPCON63025.2024.10850121
Huang, J., Wang, K., Hou, Y., & Wang, J. 2024. LW-YOLO11: A Lightweight Arbitrary-Oriented Ship Detection Method Based on Improved YOLO11. Sensors, 25(1), 65. https://doi.org/10.3390/s25010065
Juba, B., & Le, H. S. 2019. Precision-Recall versus Accuracy and the Role of Large Data Sets. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039
Kyaw, N. N., Mitra, P., & Sinha, G. R. 2024. Automated recognition of Myanmar sign language using deep learning module. International Journal of Information Technology, 16(2), 633–640. https://doi.org/10.1007/s41870-023-01680-2
Likhar, P., Bhagat, N. K., & G N, R. 2020. Deep Learning Methods for Indian Sign Language Recognition. 2020 IEEE 10th International Conference on Consumer Electronics (ICCE-Berlin), 1–6. https://doi.org/10.1109/ICCE-Berlin50680.2020.9352194
Macsik, P., Pavlovicova, J., Kajan, S., Goga, J., & Kurilova, V. 2024. Image preprocessing-based ensemble deep learning classification of diabetic retinopathy. IET Image Processing, 18(3), 807–828. https://doi.org/10.1049/ipr2.12987
Maurício, J., Domingues, I., & Bernardino, J. 2023. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Applied Sciences, 13(9), 5521. https://doi.org/10.3390/app13095521
Mousa, M. A. A., Elgohr, A. T., & Khater, H. A. 2024. A Novel Hybrid Deep Neural Network Classifier for EEG Emotional Brain Signals. International Journal of Advanced Computer Science and Applications, 15(6). https://doi.org/10.14569/IJACSA.2024.01506107
Najib, F. M. 2024. A multi-lingual sign language recognition system using machine learning. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-20165-3
Noor, T. H., Noor, A., Alharbi, A. F., Faisal, A., Alrashidi, R., Alsaedi, A. S., Alharbi, G., Alsanoosy, T., & Alsaeedi, A. 2024. Real-Time Arabic Sign Language Recognition Using a Hybrid Deep Learning Model. Sensors, 24(11). https://doi.org/10.3390/s24113683
Usama, A., Dora, M., Elhadidy, M., Khater, H., & Alkelany, O. (2021). First Person View Drone-FPV. The International Undergradu-ate Research Conference, 5(5), 437–440. https://doi.org/10.21608/iugrc.2021.246400
Vetagiri, A., Pakray, P., & Das, A. 2025. A deep dive into automated sexism detection using fine-tuned deep learning and large language models. Engineering Applications of Artificial Intelligence, 145. https://doi.org/10.1016/j.engappai.2025.110167
Wadhawan, A., & Kumar, P. 2020. Deep learning-based sign language recognition system for static signs. Neural Computing and Applications, 32(12), 7957–7968. https://doi.org/10.1007/s00521-019-04691-y
Zhang, J., Bu, X., Wang, Y., Dong, H., Zhang, Y., & Wu, H. 2024. Sign language recognition based on dual-path background erasure convolutional neural network. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-62008-z
Zhang, Y., & Jiang, X. 2024. Recent Advances on Deep Learning for Sign Language Recognition. Computer Modeling in Engineering & Sciences, 139(3), 2399–2450. https://doi.org/10.32604/cmes.2023.045731