كلية الملك عبد الله الثاني
لتكنولوجيا المعلومات
KING ABDULLAH II SCHOOL OF
INFORMATION TECHNOLOGY

Deep learning project report

Arabic sign language detection

Mohammad Al-Najjar *ID: 0207904*

Hassan Barjawi *ID:0207670*

Teacher: Tamam Al-Sarhan

Date: 4 – 1 – 2024

Contents

1.Introduction

1.1 intro

Deep learning techniques are used to enhance our comprehension of Arabic Sign Language recognition. We explain the importance of this research in its introduction, by emphasizing the need for comprehensive research in this domain, and how it can positively affect communication accessibility for people with hearing impairments. The section on related works offers an extensive examination of the current research, highlighting areas where our contributions stand out. Afterwards, the strategy outlines the selected deep learning framework and model construction, providing an explanation for why these options were made. The training, validation, and testing processes are examined in detail through our research, with the focus on evaluating the model's performance on a wide range of datasets. The final part of the report gathers the most important points made throughout, highlights their significance, and provides ideas for future studies to be carried out. The basis of this work is a solid foundation of references that includes both fundamental texts and current research that has influenced the field of deep learning in sign language recognition. These sections, when combined, give a comprehensive view of the challenges, techniques, and results in our pursuit of enhancing Arabic Sign Language recognition utilizing deep learning.

1.2 Background

According to the World Health Organization (WHO) , around 466 million people suffer from hearing loss, with 34 million of them being children. It is estimated that approximately 900 million individuals would be deafened by 2050 [1]. Hard-of-hearing people can hear to a limited extent that is undetectable by a hearing aid. Deaf persons, on the other hand, are unable to hear completely owing to head trauma, noise exposure, sickness, or a genetic problem .

Sign language is used to communicate between the deaf and the general public, and each country has its unique language. ArSL, which is used in Arabic regions, is one of these languages; it was formally launched in 2001 by the Arab Federation of the Deaf (AFOD). Sign Language relies on hand movements and gestures to communicate.

Figure 1 Comprehensive visual dataset of Arabic Sign Language (ArSL) alphabet gestures. Each image showcases a distinct hand sign corresponding to a letter from the ArSL alphabet, providing a rich resource for training convolutional neural networks (CNNs) in recognizing and interpreting sign language. This dataset aims to facilitate the development of accessible communication tools through deep learning-based sign language translation systems.

1.3 Problem definiton

The communication gap between hearing and deaf people is large, and we want to close it, but it is a long journey, therefore the best way is to study the subject from the ground up and learn the basics. To begin, we should understand the signs of the Arabic Alphabets in depth in order to lessen the hurdles that sign language learners face, but this is not an easy task for all students. Many of them will be perplexed when they study a new field, which could be problematic. As a result, we plan to create a model that detects the letter sign from Arabic Sign Language speakers.

1.4 Objective

The key goals of our research project are the creation and implementation of a YOLOv8-based deep learning model for real-time detection and recognition of Arabic sign language motions. We intend to improve the model's accuracy, speed, and robustness by experimenting with hyperparameters, data augmentation approaches, and other optimization tools. In addition, we want to improve the model's accuracy and generalization across different datasets by extending it to allow multi-class identification for various Arabic sign language gestures, ensuring real-time detection capabilities, and evaluating the model's correctness and generalization. Benchmarking existing models and providing extensive documentation for knowledge transfer are essential goals. User feedback will be gathered for iterative changes, allowing us to continuously improve our Arabic sign language detecting technology.

1.5 Research questions

In this research, we delve into the potential of deep learning to revolutionize Arabic Sign Language (ArSL) recognition. ArSL, a vital communication tool for the deaf community in Arabic-speaking regions. Our research is driven by three pivotal questions: Firstly, we explore the adaptation of deep learning models for ArSL gesture recognition. Secondly, we identify the key challenges in creating a robust ArSL detection system, encompassing gesture complexity and diverse environmental factors. Lastly, we aim to enhance the accuracy and efficiency of ArSL detection, leveraging advanced deep learning methodologies. This study is poised to contribute significantly to the field of accessible technology, bridging communication gaps for the deaf and hard of hearing community.

1.6 Challenges

Gathering data for research on Arabic Sign Language is a significant challenge due to its complex nature and how this linguistic modality works. The constraint is due to the limited access to well-annotated datasets that comprehensively represent the extensive range of Arabic sign language movements. The scarcity of datasets that are openly available for Arabic Sign Language, coupled with the intricacy of the language itself, presents a significant challenge. It is vital to consider the need for inclusivity among diverse signers while taking into account variations in their signing styles, facial expressions, and environmental circumstances when developing a robust model. The fundamental issue with this task is obtaining a dataset that has the appropriate level of detail and size, which is essential for accurately capturing the intricate subtleties and variations inherent in Arabic Sign Language. This is essential for equipping the model with the ability to adapt effectively to diverse linguistic and cultural contexts in sign language communication.

1.7 Validation methodology and software and hardware tools

In this investigation concerning, the validation methodology is methodically constructed to ensure the robustness and generalization of the employed deep learning model, in the realm of software tools, PyTorch serve as foundational framework for model development, complemented by OpenCV for testing our model in real time scenarios, and we use RoboFlow for data augmentation, training of the model was performed on an NVIDIA GeForce RTX 3090 GPU

2. Related work

Many researchers have been studying sign language recognition using various approaches since 1990 [2] [3] [4] [5] .In this section, we will discuss the related literature, which comprises several works and methods developed over the years. Researchers, led by Tamura and their team [6], had a theory. They thought that a sign word is made up of a series of time-based units called cheremes. These cheremes include information about the hand's shape, movement, and where the hand is positioned. To understand these handshape, movement, and location features, the researchers represented them in three dimensions (3D). Then, they transformed these 3D features into two-dimensional (2D) image features. Afterward, they used these 2D features to categorize or classify motion images of sign language.

Researchers led by Keskin and their team [7] developed lifelike 3D models of hands. These models broke down the hand into two different parts. To make these models learn and understand, they trained Random Decision Forests (RDFs), a type of algorithm. The RDFs were used to classify each pixel in an image, assigning it to a specific part of the hand. They then applied a local mode-finding algorithm to figure out where the joints of the hand skeleton are located. The team also introduced a model based on Support Vector Machine (SVM) to recognize Arabic Sign Language (ASL) digits using this method. Impressively, their system achieved a high rate of accuracy in recognizing live depth images in real-time. In another study, Nandy and colleagues [8] established a video database containing different signs from the Indian sign language. To analyze and understand these signs, they employed a direction histogram a tool that is robust to changes in lighting and orientation—as the features for classification. They utilized two distinct methods for recognition: the Euclidean distance and the K-nearest neighbor metrics. These techniques helped them effectively identify and categorize various signs in the Indian sign language based on the provided features.

In a study conducted by Mehdi and colleagues [9], they employed a 7-sensor glove manufactured by the 5DT Company to record data about hand movements. This data was then processed using artificial neural networks (ANN) to recognize sign gestures, achieving an accuracy of 88%. Following a similar path, Lopez-Noriega et al. [10] not only adopted the same approach but also created a user-friendly graphical interface using '.NET'. They used a Hidden Markov Model (HMM) based model, which demonstrated effectiveness in real-time sign language recognition tasks.

Another noteworthy approach was taken by Starner and team [11].They utilized images of gloves as input for their Hidden Markov Model (HMM)-based system. This method involved using colored gloves to capture essential features like hand shape, orientation, and trajectory. Their proposed recognition system based on HMM proved highly accurate, especially in recognizing sentence-level American Sign Language (ASL), achieving impressive word accuracy results.

In the study by Hienz and team [12], they used colored cotton gloves to simplify the process of extracting features. Their method involved converting sequences of videos into feature vectors, which were then processed by a Hidden Markov Model (HMM) for classification, achieving accuracy values between 92% and 94%. Similarly, Grobel et al. [13] and Parcheta et al. [14]employed comparable techniques. Despite the success of these earlier approaches in achieving high accuracy, they faced practical limitations. They required users to wear gloves and were confined to specific environments, making them less suitable for everyday use. Moreover, many of these methods were user-dependent, necessitating individual training for each user, which was considered impractical and unnatural. In response to these challenges, Youssif et al. [15] sought a more generalized approach and proposed an HMM-based model that didn't rely on user-specific training or the use of gloves. However, their model encountered a trade-off, resulting in a lower accuracy rate of 82%.

Convolutional Neural Networks (CNN) are widely recognized for their effectiveness in image recognition and classification, and this extends to their application in Sign Language Recognition (SLR). Masood and colleagues [16] delved into ASL character recognition, presenting a CNN model that demonstrated remarkable performance—achieving an overall accuracy of 96% across a dataset containing 2,524 ASL gesture images. In parallel, Wadhawan et al. [17], Bheda et al. [18], and Tao et al. [19] also harnessed the power of CNNs for sign language recognition. Wadhawan et al. achieved an impressive accuracy of 99% when classifying signs from different languages. Bheda et al. achieved 82.5% accuracy, while Tao et al. reached a remarkable 100% accuracy in their respective studies. These findings underscore the efficacy of CNNs in the domain of sign language recognition across different languages and alphabets.

Convolutional Neural Networks (CNNs) typically work on a frame-by-frame basis, making them well-suited for image recognition. However, when paired with Recurrent Neural Networks (RNNs), CNNs can extend their capabilities to retain information over time, which is particularly beneficial for video analysis. This synergy is advantageous for recognizing dynamic signs accurately. Yang and colleagues [20] introduced an innovative continuous sign language recognition method, capitalizing on the integration of CNN and Long Short-Term Memory (LSTM). Their experiments, conducted on a dataset they created, showcased impressive accuracies.

In contrast, 3D Convolutional Neural Network (3D-CNN) models, as opposed to 2D-CNNs, inherently incorporate temporal information. They can capture multi-frames of a video simultaneously, eliminating the need for an additional RNN phase. Huang et al. [21] and Al-Hammadi et al. [22] embraced this approach in their proposed models.

The current study follows the CNN-RNN approach, utilizing double CNNs as feature extractors. For the RNN component, Bi-directional Long Short-Term Memory (BiLSTM) layers are employed. The BiLSTM layers play a crucial role in identifying complex sequences in videos, addressing potential conflicts between different sign classes."

3. Methodology

Model Architecture:

We use YOLOv8 model, YOLOv8 is the latest iteration in the YOLO series of real-time object detectors, offering cutting-edge performance in terms of accuracy and speed. Building upon the advancements of previous YOLO versions, YOLOv8 introduces new features and optimizations that make it an ideal choice for various object detection tasks in a wide range of applications [23] [24].

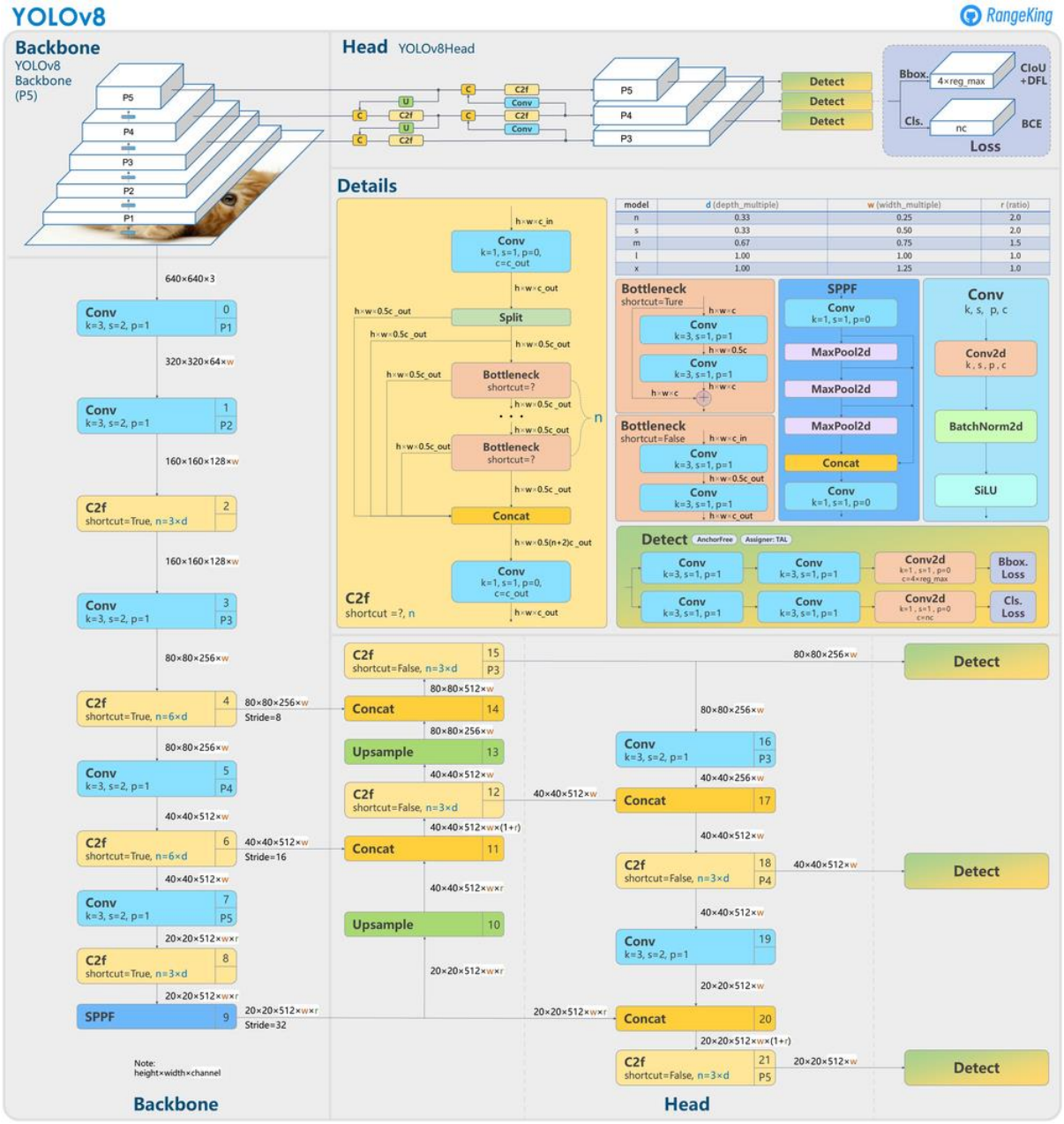


Figure 2 Architectural diagram of YOLOv8 showcasing the model's detailed composition. The backbone consists of successive convolutional and C2F layers, enabling robust feature extraction across multiple scales (P1-P5). The head, integrating YOLOv8Head, processes these features through convolutional, bottleneck, and upsampling layers, culminating in object detection across different scales. Loss functions are applied at the output for bounding box, class, and objectness predictions, contributing to the model's precision in real-time object detection tasks [24].

The following key advancements underscore its architectural prowess:

- **Advanced Backbone and Neck Architectures:** YOLOv8 employs state-of-the-art backbone and neck architectures, resulting in improved feature extraction and object detection performance.
- **Anchor-free Split Ultralytics Head:** YOLOv8 adopts an anchor-free split Ultralytics head, which contributes to better accuracy and a more efficient detection process compared to anchor-based approaches.
- **Optimized Accuracy-Speed Tradeoff:** With a focus on maintaining an optimal balance between accuracy and speed, YOLOv8 is suitable for real-time object detection tasks in diverse application areas.
- **Variety of Pre-trained Models:** YOLOv8 offers a range of pre-trained models to cater to various tasks and performance requirements, making it easier to find the right model for your specific use case.

Functional Mechanism:

Operating seamlessly, YOLOv8 processes input images through its CNN architecture. The network adeptly learns to recognize and precisely locate objects within images by predicting bounding boxes along with their associated class labels. Its prowess lies in its real-time object detection capabilities, owing to an intricately efficient architecture and meticulous design optimization.

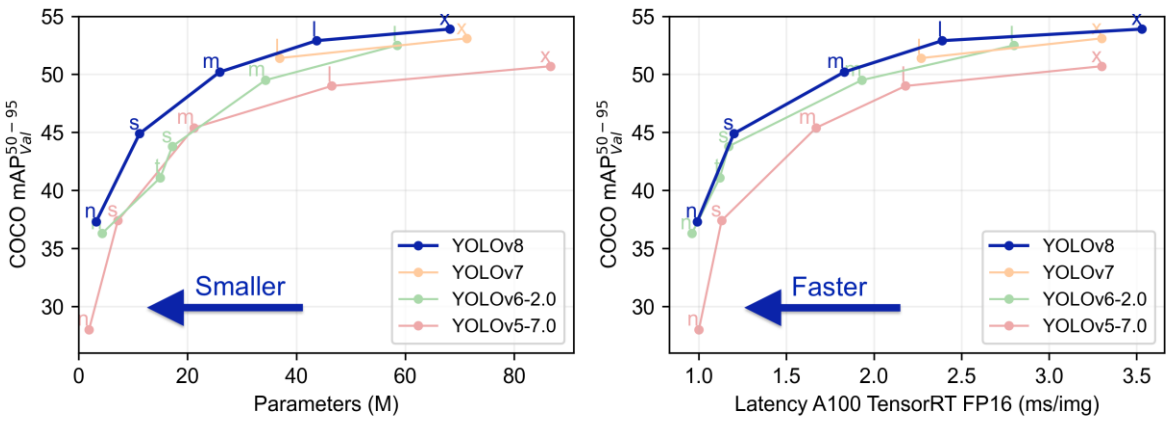


Figure 3 Performance comparison of YOLO versions in terms of accuracy and speed. The left graph shows the mean Average Precision (mAP) on the COCO dataset against the model parameters, illustrating YOLOv8's superior accuracy with fewer parameters, indicating a smaller model size. The right graph compares inference speed (latency) with mAP, demonstrating YOLOv8's faster performance on an NVIDIA A100 using TensorRT FP16, emphasizing its efficiency in real-time applications [24].

4.Experiments

4.1 Data Preparation:

we use a public dataset for arabic sign language , the dataset is composed of 290 images for testing set, 4651 images for training set, 891 images for validation set, for a total of 5832 images with each image having the size of 416 × 416 pixels, these images was taken in different environments using a cell phone camera with different backgrounds and different hand angles, this is the unaugmented version of Arabic sign language dataset [25] .

and we see that to enhance the robustness of our model, our dataset need to be underwent on several preprocessing methods. Since all input images in the dataset were of size 416 x 416 pixels, image resizing was required, so we resized the images to 640 x 640 pixels. Data augmentation was performed to increase the dataset's size and diversity, two types of modifications were applied:

-Image rotation is applied for data augmentation to increase the model's robustness by exposing it to variations in orientation

-Bounding box noise is employed to improve the model's resilience and enhance its generalization capabilities.

4.2 Training Procedure:

we cloned the yolov8 repository Hyperparameter tuning was conducted and evaluating results. as shown in Table. Training of the model was performed on an NVIDIA GeForce RTX 3090 GPU

Hyperparameter	Value
epochs	25
Batch size	16
Learning rate	0.01
optimizer	Adam

4.3 Evaluation Metrics

1-Mean Average Precision: Mean Average Precision (mAP), computed using equation 1, is a widely accepted performance metric for object detection models. mAP is calculated by taking the mean of Average Precision (AP) for each of the 'n' classes. AP for each class 'k' is determined by calculating the area under the precision-recall curve. mAP provides a single score that considers Recall, Precision, and Intersection over Union (IoU), eliminating bias in performance measurement [26] .

The equation for Average Precision (AP) is as follows:

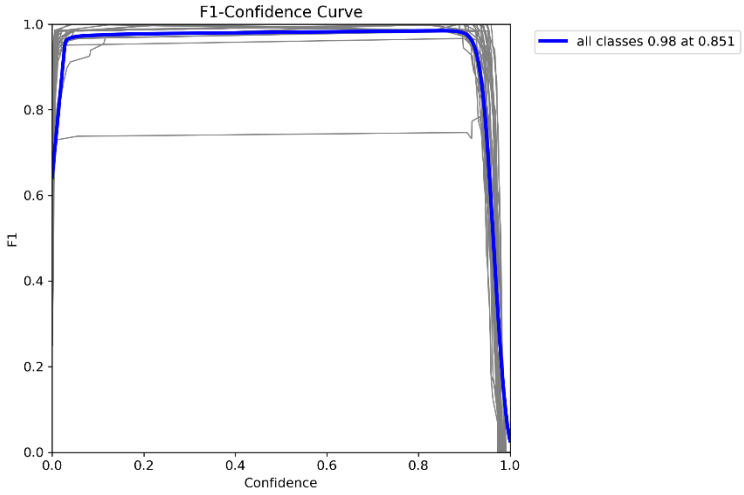
$$AP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

Class	mAP @ 50	mAP @ 50 - 95
All	0.984	0.896
ALIF	0.995	0.815
BAA	0.995	0.893
TA	0.995	0.904
THA	0.995	0.908
JEEM	0.995	0.897
HAA	0.995	0.869
KHAA	0.95	0.836
DELL	0.995	0.906
DHELL	0.995	0.91
RAA	0.995	0.913
ZAY	0.995	0.915
SEEN	0.995	0.93
SHEEN	0.995	0.913
SAD	0.995	0.874
DAD	0.995	0.956
TAA	0.995	0.942
DHAA	0.995	0.948
AYN	0.995	0.894
GHAYN	0.977	0.911
FAA	0.995	0.918
QAAF	0.742	0.647
KAAF	0.995	0.935
LAAM	0.995	0.93
MEEM	0.995	0.902
NOON	0.995	0.899
HA	0.995	0.886
WAW	0.995	0.933
YA	0.995	0.914

Figure 4 MEAN AVERAGE PRECISION (MAP) METRICS FOR OUR MODEL

2-F1 score

The F1 score is a metric that combines precision and recall to evaluate the performance of a binary classification model. It provides a balanced measure of accuracy, especially when the cost of false positives and false negatives differs. The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall [27] and we achive 0.98 for all classes

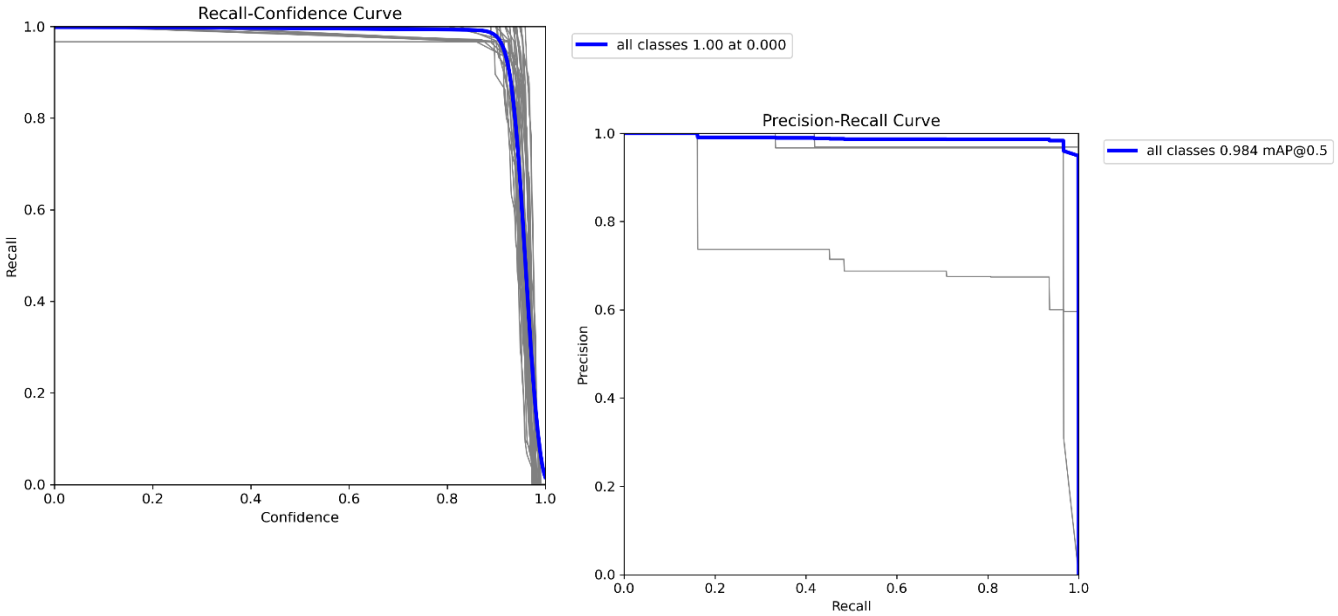


$$2\ x\ \frac{precision\ x\ recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

3-recall

Recall measures a model's ability to capture all positive instances. It is especially important when missing positive instances has significant consequences and it ranges from 0 to 1, with 1 indicating perfect recall [28], and we achive 0.984 mAP @ 0.5

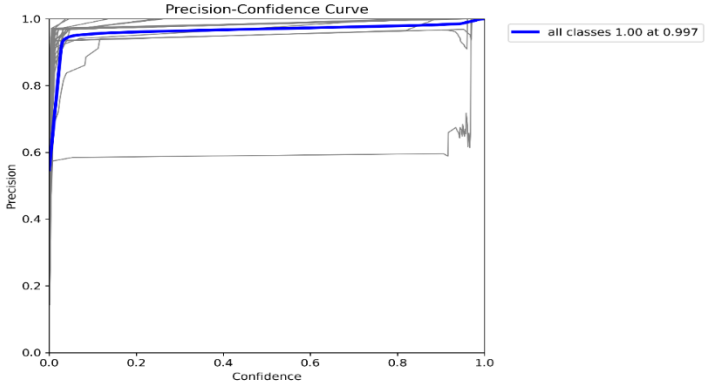
$$recall = \frac{TP}{TP+FN}$$



4- Precision

Precision is a metric used to evaluate the accuracy of positive predictions made by a binary classification model, Precision is particularly important when minimizing false positives is crucial , And precision ranges from 0 to 1, with 1 indicating perfect precision, and we achive 0.997

$$\text{Precision} = \frac{TP}{TP+FP}$$



5. Conclusion

In conclusion, this research has ventured into the burgeoning field of Arabic Sign Language (ArSL) recognition through the lens of deep learning, addressing the communication barriers faced by the deaf community. We have embarked on a journey to bridge the gap between the hearing and the deaf by delving into the subtleties of ArSL, beginning with the nuanced signs of the Arabic alphabet.

Our project's cornerstone is the development of a YOLOv8-based deep learning model tailored for the real-time detection and recognition of ArSL gestures. Rigorous experimentation with hyperparameters, data augmentation, and optimization techniques has enhanced the model's accuracy, speed, and robustness. The model has been fine-tuned for multi-class identification, thereby expanding its applicability across a diverse array of ArSL gestures. Critical to our endeavor was the evaluation of the model's performance, ensuring its capability to operate in real-time while maintaining high accuracy and generalization across various datasets.

The benchmarks established against existing models have set a new standard for the efficacy of ArSL recognition systems. Comprehensive documentation and knowledge dissemination have been prioritized, ensuring that the insights and findings from this research can serve as a foundation for future work in the field. User feedback has been integral, fostering a cycle of continuous refinement and adaptation that keeps the user's needs at the forefront.

As we look to the future, the potential for further advancements in ArSL recognition is vast. The aspirations for subsequent studies include enhancing the model's contextual understanding, expanding the dataset to cover a broader spectrum of dialects and colloquial variations, and exploring the integration of this technology into a range of assistive devices. The ultimate aim is to not only improve communication accessibility for the deaf community in Arabic-speaking regions but also to contribute meaningfully to the inclusivity and diversity of global communication tools. This research is a step toward a world where language barriers are surmounted, and understanding is within everyone's reach.

References

[1] W. H. O. (WHO), "New WHO-ITU standard aims to prevent hearing loss among 1.1 billion young people," 2019. [Online]. Available: https://www.who.int/news/item/12-02-2019-new-who-itu-standard-aims-to-prevent-hearing-loss-among-1.1-billion-young-people#:~:text=Over%205%25%20of%20the%20world's,%2D%20and%20middle%2Dincome%20countries..

[2] B. H. B. W. B. M. Abdulazeem Y, "Human action recognition based on transfer learning approach," IEEE Access 9:82058–82069, 2021.

[3] H. B. B. R. Cooper H, "Sign language recognition in visual analysis of humans," 2011.

[4] P. A. Starner T, " Real-time american sign language recognition from video using hidden markov models in Motion-based recognition," 1997.

[5] W. J. P. A. Starner T, " Real-time american sign language recognition using desk and wearable computer based video," IEEE Trans Pattern Anal Mach Intell 20(12):1371–1375, 1998.

[6] K. S. Tamura S, "Recognition of sign language motion images," 1988.

[7] K. F. K. Y. A. L. Keskin C, "Real time hand pose estimation using depth sensors in consumer depth cameras for computer vision," 2013.

[8] P. J. M. S. C. P. N. G. Nandy A, "Recognition of isolated indian sign language gesture in real time. In: International conference on business administration and information processing," 2010.

[9] K. Y. Mehdi SA, " Sign language recognition using sensor gloves. In: Proceedings of the 9th international conference on neural information processing," ICONIP'02, vol 5. IEEE, pp 2204– 2206, 2002.

[10] F. ´. a.-V. M. U.-C. V. Lopez-Noriega JE, "Glove-based sign language recogni- ´ tion solution to assist communication for deaf users.," 2014.

[11] S. TE, " Visual recognition of american sign language using hidden markov models.," 1995.

[12] B. B. K. K. Hienz H, "Hmm-based continuous sign language recognition using stochastic grammars.," 1999.

[13] A. M. Grobel K, "Isolated sign language recognition using hidden markov models," 1997.

[14] M.-H. C. Parcheta Z, "Sign language gesture recognition using hmm. In: Iberian conference on pattern recognition and image analysis. Springer, pp 419–426," 2017.

[15] A. A. A. H. Youssif A, "Arabic sign language (arsl) recognition system using hmm.," International Journal of Advanced Computer Science and Applications (IJACSA) 2(11), 2011.

[16] T. H. S. A. Masood S, "American sign language character recognition using convolution neural network in Smart Computing and Informatics.," Springer, pp 403–412, 2018.

[17] K. P. Wadhawan A, " Deep learning-based sign language recognition system for static signs," Neural Comput Applic, 1–12, 2020.

[18] R. D. Bheda V, "Using deep convolutional networks for gesture recognition in american sign language.," arXiv:1710.06836, 2017.

[19] L. M. Y. Z. Tao W, "American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion.," Eng Appl Artif Intell 76:202–213, 2018.

[20] Z. Q. Yang S, "Continuous chinese sign language recognition with cnn-lstm.," In: Ninth international conference on digital image processing (ICDIP 2017). (International Society for Optics and Photonics), vol 10420, p 104200F, 2017.

[21] Z. W. L. H. L. W. Huang J, "Sign language recognition using 3d convolutional neural networks.," In: 2015 IEEE international conference on multimedia and expo (ICME). IEEE, pp 1–6, 2015.

[22] A.-H. M. e. al, " Hand gesture recognition for sign language using 3dcnn.," IEEE Access 8:79491–79509, 2020.

[23] J. D. S. G. R. &. F. A. Redmon, "You only look once: Unified, real-time object detection," 2016.

[24] Ultralytics, "Ultralytics YOLOv8 Docs," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[25] Kaggle, "Arabic Sign Language ArSL dataset," [Online]. Available: https://www.kaggle.com/datasets/sabribelmadoui/arabic-sign-language-unaugmented-dataset.

[26] KILI, "Mean Average Precision," [Online]. Available: https://kili-technology.com/data-labeling/machine-learning/mean-average-precision-map-a-complete-guide.

[27] Wikipedia, "F-score," 1992. [Online]. Available: https://en.wikipedia.org/wiki/F-score.

[28] iguazio, "Recall," [Online]. Available: https://www.iguazio.com/glossary/recall/#:~:text=Recall%2C%20also%20known%20as%20the,total%20samples%20for%20that%20class..

[29] L. D. S. K. P. J. &. S. B. Pigou, " Sign language recognition using convolutional neural networks," European Conference on Computer Vision, Springer, Cham, 2014.