INTJOS
INTERNATIONAL JOURNAL OF SUSTAINABILITY
ULUSLARARASI SÜRDÜRÜLEBİLİRLİK DERGİSİ
intjos.com
Cilt(Volume): 3 | Sayı (Number): 1 | Yıl (Year): 2025 | e-ISSN: 2980-1338

INTERNATIONAL JOURNAL OF SUSTAINABILITY
DOI: https://doi.org/10.5281/zenodo.17624064
INTJOS 2025; 3 (1) :29-44

Sign Language Recognition Modeling With Deep Learning Method

Derin Öğrenme Yöntemiyle İşaret Dili Tanıma Modellemesi

● Mehmet AŞIROĞLU
Üsküdar Amerikan Lisesi
mehmet.asiroglu07@gmail.com
https://orcid.org/0009-0006-1883-2245

● Atınç YILMAZ
İstanbul Beykent Üniversitesi,
Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye
atincyilmaz@beykent.edu.tr
https://orcid.org/0000-0003-0038-7519

● Oğuz Emre UZMAN
İstanbul Beykent Üniversitesi,
Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye
oguzemreuzm@hotmail.com
https://orcid.org/0009-0000-9736-5265

Kaynak gösterimi için:
AŞIROĞLU M. & YILMAZ A..& UZMAN O.E.. (2025). Sign Language Recognition Modeling With Deep Learning Method; International Journal of Sustainability -INTJOS, c.3 **s.1 ISSN: 2980-1338**

ÖZET

İşaret dili, işitme veya konuşma engelli bireyler için temel bir iletişim aracıdır; ancak otomatik tanıma teknolojileri hâlâ sınırlıdır. Bu çalışma, işaret hareketlerinden oluşan çeşitli örnekler sunan Word-Level American Sign Language (WLASL) veri kümesi üzerinde değerlendirilen işaret tanıma çerçevesi sunmaktadır. Önerilen yaklaşım, yalnızca standart kamera girişleri kullanarak mekânsal el konfigürasyonlarını doğru biçimde modellemeye olanak tanıyan ResNet50 tabanlı evrişimsel özellik çıkarımı ile MediaPipe çerçevesinden elde edilen el işaret noktası temsillerini birleştirmektedir. Model eğitimi, WLASL veri kümesindeki 15 yaygın işaret sınıfı kullanılarak Create ML ortamında gerçekleştirilmiştir. Deneysel sonuçlar, yaklaşık %94 doğruluk oranı ile sınıflandırma performansı elde edildiğini ve yanıt sürelerinin etkileşimli uygulamalar için uygun seviyede olduğunu göstermektedir. Karışıklık matrisi analizi, modelin farklı işaretleri tanımadaki güçlü yönlerini ortaya koyarken, görsel olarak benzer işaretler arasındaki zorlukları da vurgulamaktadır. Teknik başarımının ötesinde, hafif mimari enerji tüketimini en aza indirir ve özel donanım gerektirmez; bu yönüyle çevresel açıdan sürdürülebilir yapay zekâ uygulamalarını destekler. Ayrıca, işitme engelli bireylere yönelik iletişim araçlarının geliştirilmesini destekleyerek, önerilen çerçeve sosyal kapsayıcılık hedeflerine katkıda bulunmakta ve eğitim ile sağlık alanlarındaki daha geniş sürdürülebilir kalkınma amaçlarıyla uyum göstermektedir.

Anahtar Kelimeler: *Yapay Zeka, İşaret Dili, Derin Öğrenme, İletişimde Süreklilik*

ABSTRACT

Sign language is a critical medium of communication for individuals with hearing or speech impairments, yet automated recognition technologies remain limited for many regional variants. This study presents a sign recognition framework evaluated on the Word-Level American Sign Language (WLASL) dataset, which provides diverse samples of isolated sign gestures. The proposed approach combines ResNet50-based convolutional feature extraction with hand landmark representations obtained through the MediaPipe framework, enabling accurate modeling of spatial hand configurations using only conventional camera inputs. Model training was performed via Create ML environment on 15 frequently used gesture classes from the WLASL dataset. Experimental results demonstrate an overall classification accuracy of approximately 94%, with response times suitable for interactive applications. Confusion matrix analysis highlights both the strengths of the model in recognizing distinct gestures and the challenges posed by visually similar signs. In addition to its technical performance, the lightweight design minimizes energy consumption and does not require specialized hardware, supporting environmentally sustainable AI practices. Furthermore, by supporting the development of communication tools for hearing-impaired individuals, the framework contributes to social inclusion objectives and aligns with broader sustainable development goals in education and healthcare.

Keywords: *Artificial Intelligence, Sign Language, Deep Learning, Sustainability in Communication.*

INTRODUCTION

It is generally accepted that inclusive communication is necessary to guarantee that people can engage in social life, healthcare, and education on an equal basis. The primary means of expressing thoughts, intentions, and even emotional states for those who are deaf or hard of hearing is sign language. It conveys meaning through a combination of body posture, facial expressions, and hand gestures. However, the general public hardly ever understands this mode of communication. Consequently, obstacles to communication continue to exist. These obstacles are not insignificant; they frequently interfere with day-to-day activities, restricting work opportunities, making it challenging to obtain healthcare, and occasionally resulting in social isolation.

The need for more accessible communication tools has gained prominence in recent years. Researchers are now focusing on automated sign language recognition (SLR) systems as a result of the demand. With the growth of computer vision and deep learning, it is now possible to translate visual sign inputs into spoken or written language. Studies conducted on widely known sign languages—such as American Sign Language (ASL), British Sign Language (BSL), and South African Sign Language—show that these systems can reach high accuracy levels and be used in practical settings (Takyi et al., 2025). Neural network models, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), play an important role here because they are capable of learning both spatial and temporal aspects of gestures.

Even with these advances, some challenges remain unsolved. Variations in signer appearance, overlapping gestures, and problems caused by lighting or camera movement still reduce system reliability. Another point is that most studies work with widely standardized sign languages. Local or regional sign variants, which are common in real communities, are often ignored. In addition to accuracy of recognition, there are additional practical concerns to account for. Most systems demand expensive hardware or cloud computer infrastructure, which makes them impractical to implement within schools, public offices, or rural health clinics. Additionally, power usage by high-powered AI models has sparked concern over such systems' feasibility. To counteract such issues, this paper advocates for a lightweight system that has been deployed with the Word-Level American Sign Language (WLASL) dataset. Efficiency and reduced reliance on hardware are made central, in line with sustainable design practices. The approach hopes to provide an example of how inclusive sign languages technology can be efficient as well as environmentally sustainable.

1. LITERATURE REVIEW

Sign language recognition (SLR) research has made great progress in the past decade. Such advances are due to advances in machine learning algorithms, growth in annotated datasets, and an increasing societal concern with inclusive communication tools. In this domain, three main lines of methodology have appeared: systems that are grounded on physical sensors, ones that are based exclusively on vision, and hybrid ones that exploit a variety of data modalities to boost robustness and accuracy.

Sensor-based methods form the earliest generation of SLR technologies. These systems usually use data gloves, accelerometers, or inertial measurement units (IMUs) to get very detailed information about how hands and fingers move (Hasan & Mishra, 2012). Even though these kinds of setups can give very accurate motion data, they need special hardware, which makes them less scalable and less useful in everyday situations. Foundational surveys, such as the review by Walde et al. (2017), delineated initial taxonomies of gesture modeling techniques and classification strategies (Walde et al., 2017). These early studies had an impact on later research paths and helped set the standard for comparing newer methods.

With the rapid growth of computer vision, focus in the field gradually shifted toward camera-based recognition. Such approaches circumvent spatial sensor requirements and process video or image inputs fed from conventional cameras rather. Veale et al. (1998) multilingual “Zardoz” system was one of the earliest significant systems to fall into this category, combining facial expressions with linguistic processing to serve sign translation ends (Veale et a., 1998). Following experiments employed convolutional neural networks (CNNs) to procure mobile and real-time deployments; Sutjiadi (2020), for example, instantiated an Android OS-based ASL finger-spelling recognizer that had been trained using general image dataset (Sutjiadi, 2020). Most recently, spatial and temporal dynamics of sign gestures have been addressed by three-dimensional CNNs at once to significantly boost interpretation of complex motion (Renjith et al., 2024). The hybrid methods that combine skeletal landmark data with raw video attributes have themselves grown popular as one way to minimize environmental variability. The recent paper by El-Alfy et al. (2022) presented an excellent review of this type, underlining the broader trend to migrate from sensor-based to visionbased approaches and noting an emerging emphasis upon multimodal fusion (El-Alfy et al., 2022).

Even with these advancements, there remains a significant lack of research on Turkish Sign Language (TSL) when compared to the extensive studies conducted on other internationally recognised sign languages that are commonly utilised. Notable attempts include the detection of TSL signs by Aksoy et al. (2021), who combined deep learning and image processing methods and reported excellent results on a controlled dataset (Aksoy et al., 2021). To demonstrate that vision-based methods can be effectively modified for regional languages, Güney and Erkuş (2022) created a CNN-based real-time recognition system specifically designed for TSL (Güney et al., 2022). The YOLOv8 architecture was used in a more recent object detection framework by Karakan and Oğuz (2025) for TSL letter and number recognition in live video streams. The framework achieved competitive results, including 90.7% stability, 85.8% mean average precision, and 81.4% recall (Karakan et al., 2025).

A recurring issue in the literature is striking a balance between practical usability and high recognition accuracy. It is challenging to implement many high-performing models in public institutions, educational institutions, or rural healthcare settings because they require resource-intensive GPUs or specialized sensors. Furthermore, the majority of studies focus more on recognition accuracy than on latency, energy efficiency, and environmental sustainability—factors that are becoming more and more important for real-world deployment. By presenting a sign recognition framework, the current study allays these worries. This strategy seeks to close the gap between scalable and ecologically friendly assistive solutions and experimental prototypes.

3. METHODOLOGY

3.1 Theoretical Background

To autonomously identify sign language, it is necessary to acquire knowledge of the spatial and temporal dynamics of human gestures. Deeper neural network architectures are suitable for this job. Russell and Norvig (2020) gave a full explanation of the theoretical bases for this. They also laid the groundwork for most of today's computer vision systems (Russell et al., 2020). Convolutional Neural Networks (CNNs) have been especially useful for problems involving visual recognition in this context. They can learn small differences in visual data because they can process with local receptive fields, share weights between layers, and build a hierarchy of feature maps. This skill is very important for understanding sign language because the ability to distinguish small changes in hand, finger, and relative positions is essential for accurate grouping of similar gestures.

The Residual Network (ResNet) setup is an improvement on the basic CNN structure. It fixes the common problem of vanishing gradients, which happens in very deep networks, by using shortcuts or residual connections. These links make identity mappings, which help the model learn by making the gradients move smoothly. This enables the construction of deeper models without compromising efficiency. The ResNet50 variant is based on a convolutional architecture and has fifty layers. Many people know that this kind of design can store a lot of data and process it quickly. Das et al. (2024) say that ResNet50 is a great choice for apps that need to track movements in real time because all of its parts work together.

Some features of this CNN backbone can be utilized. Using landmark-based extraction methods like those from MediaPipe is another way to learn more about how the hands move. In two or three dimensions, MediaPipe can find 21 important places on each hand. These points show where the tips and joints of the fingers are. The skeletal form can handle changes in lighting or being partially blocked and still keep the fingers' positions in space. Combining pixel features from CNN with these famous places supports a hybrid approach that uses both low-level visual clues and higher-level structural descriptors. The method makes classification more accurate without adding much latency, which is useful for applications that need to read signs.

3.2. Deep Learning

Deep learning has become a key method in AI because it can find important features in raw data without needing to be manually engineered (Alhijaj et al., 2023). Using multilayered neural network structures, these models can slowly learn hierarchical representations. This skill makes it much easier to deal with difficult things like processing text, audio, and images. This skill is very important for understanding sign language because even small changes in how the hands are positioned, how they move, or how the fingers are positioned can change how a sign is understood.

In the context established by this study, deep learning functions as the principal analytical component. The system does not just look at frames one at a time; it looks at whole video sequences to find the spatiotemporal dependencies between movements that happen one after the other. This time dimension is very important in sign languages because the meaning comes from how the gestures move, not from still images.

Another good thing about the proposed method is that it uses three-dimensional coordinates (x, y, z) that match up with anatomical landmarks like the wrists, elbows, shoulders, and fingertips. The model's job is easier when it uses these coordinates instead of raw pixel data. This helps it focus on the geometric relationships that are most important for classification. Data changes into more and more abstract forms as it moves through different layers of the network. This process keeps the model strong even when the lighting, background, or the way the signer signs changes. Training in steps also helps reduce errors in classification.

3.3. ResNet50 Architecture

ResNet50 is a well-known standard architecture in computer vision because it balances network depth and speed well. It learns residual functions by using identity shortcut connections instead of direct mappings. It has 50 layers that are grouped into residual blocks (Das et al., 2024). This residual design helps with the problem of the gradient disappearing, which is common in deep networks. This makes training more stable and allows the use of much deeper architectures in a meaningful way.

ResNet50 has an initial convolutional layer, a max pooling stage, and four more residual stages. The number of feature dimensions gradually increases from 64 to 128, 256, and finally 512 channels during these phases. In the bottleneck configuration of each residual block, a 1×1 convolution lowers the number of dimensions, a 3×3 convolution pulls out spatial features, and a final 1×1 convolution raises the number of dimensions again. This design is great for tasks that need accurate but scalable feature extraction because it strikes a good balance between how well it works and how well it represents things.

This hierarchical structure makes it much easier to understand sign language. ResNet50's multiscale features can accurately capture both small finger movements and bigger arm movements. This is because gestures often combine the two. The network was started with pretrained weights from big datasets like ImageNet. These datasets provide basic visual elements like edges, contours, and textures that go well with sign language data. Fine-tuning the top layers on the WLASL dataset helps cut down on training time, overfitting, and makes the model more accurate in different recording situations. It also helps the network adjust to how the task changes.

3.4. Rationale for Model Selection

ResNet50 was chosen as the framework for this study because it performs well with both empirical data and real-world application requirements. ResNet50 is one of the deep convolutional architectures that produces successful results. Furthermore, it utilizes residual connections to solve the vanishing gradient problem, a common problem in deep learning. This feature makes it particularly suitable for situations where datasets are relatively small, such as those in sign languages that are not sufficiently represented by examples.

Its compatibility with transfer learning is equally important. The network inherits low-level visual features and requires less training data to achieve competitive performance by initializing with pretrained weights from large datasets like ImageNet. Given the scarcity of annotated sign language corpora, this method reduces the chance of overfitting while speeding up development cycles.

The system incorporates landmark-based hand representations that are extracted using the MediaPipe framework to supplement this backbone. MediaPipe maintains the geometric relationships between joints and fingertips while identifying 21 important landmarks per hand. By using these structured coordinates instead of pixel-level images, computational overhead is greatly decreased while the essential spatial information required for precise recognition is preserved. For deployment on consumer-grade devices with constrained processing power and energy resources, this efficiency is especially beneficial.

The Create ML environment was used to implement the entire processing pipeline, from preprocessing to training, validation, and deployment (Apple Inc., 2024). The ResNet50 backbone receives landmark coordinates from WLASL video sequences and uses them to build hierarchical feature representations through bottleneck and residual layers. Classification is then achieved through fully connected layers with softmax activation. As a result, Figure 1 shows a model that is optimized for performance on the macOS and iOS platforms.

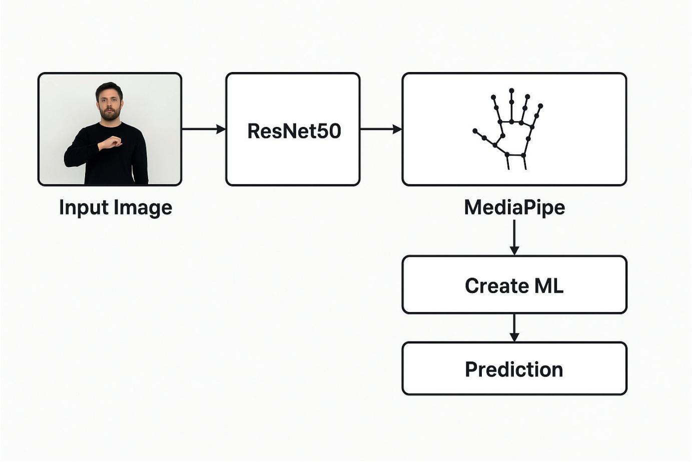

Figure 1. The Proposed System Architecture Workflow

A final consideration guiding the architectural choice was sustainability. The suggested framework's low hardware requirements and energy-conscious design complement more general guidelines for developing AI in an environmentally responsible manner. The method provides a scalable route for inclusive communication technologies by preserving predictive accuracy while consuming the fewest resources possible. It is especially appropriate for educational and public service settings where computational and energy resources are frequently limited.

The following is a summary of this methodology's contributions:

- A fusion of residual network hierarchies with skeletal landmarks, striking a balance between recognition accuracy and computational demands.
- An effective transfer learning adaptation, fine-tuning pretrained weights to address the scarcity of annotated data in regional sign language.
- A deployment-oriented design optimized for execution on standard hardware, promoting accessibility in diverse environments.
- Integration of sustainability principles by reducing energy consumption and hardware requirements, thus enabling environmentally responsible AI solutions.

In this study, deep learning techniques are combined with compact skeletal representations to address persistent challenges commonly observed in sign language recognition research. The primary obstacles include the restricted accessibility of annotated training data and the computational limitations imposed by hardware with limited resources. In addition to addressing these technical issues, the suggested method supports a more general goal: creating recognition frameworks that support social inclusion while preserving energy efficiency and reducing environmental impact.

4. EXPERIMENTAL STUDIES

The empirical studies carried out to evaluate the effectiveness and viability of the suggested sign language recognition framework are presented in this section. The evaluation covers a number of important topics, including network architecture, feature extraction techniques, preprocessing strategies, dataset characteristics, and training protocols. Apart from standard metrics for recognition accuracy, the experiments were designed to assess computational efficiency and the system's suitability for deployment, which are crucial considerations when transferring assistive technologies from research prototypes to everyday use.

4.1. Dataset and Preprocessing Process

A dataset that was first made public during the Kaggle Sign Language Recognition competition was used in the experimental phase (Kaggle, 2025). This resource comes from the Word-Level American Sign Language (WLASL) corpus, which has more than 40,000 annotated examples of 30 common manual gestures. Instead of storing whole pixel-level video frames like most datasets do, this corpus encodes each gesture using landmark coordinates that cover the hands, face, and upper body. This kind of representation is small in terms of computation and keeps semantically important structural information, which makes it easier to model gesture dynamics well and uses less memory.

For this study, we chose 15 gesture classes that are often used in everyday communication to replicate interactions. Both static signs (like numbers or commonly used words) and dynamic movements (like greetings or affirmative responses) were included in the chosen gestures. Instead of being randomly selected, these classes were specifically chosen to represent gestures that are frequently used in daily interactions, distinct interclass differences, and a range of temporal characteristics that are appropriate for assessing both isolated and transitional sign recognition.

The dataset was divided in order to guarantee a thorough assessment. In order to closely resemble real-world deployment conditions where the system must generalize to previously unseen signers, individuals who were part of the training subset were excluded from the testing subset. Stratified sampling was used to ensure balanced class distributions across all subsets, and the data were split 80-10-10 for training, validation, and testing.

To improve robustness, several preprocessing steps were used before model training. To adjust for variations in camera distance and scale, landmark coordinates were normalized in relation to frame dimensions. Then, to reduce slight jitter in keypoint detection—a modification that is especially helpful for dynamic gestures—temporal smoothing was used. Additionally, data augmentation techniques, such as random horizontal flipping and mild temporal scaling, were used to boost intra-class variability and strengthen the system's resistance to changes in the environment. All of these preprocessing steps combined produced a dataset that was ideal for precise and effective recognition in the suggested framework.

4.2. Preprocessing and Feature Engineering

The goal of designing the preprocessing and feature engineering pipeline was to guarantee consistent performance across various signers and under various recording conditions. Data in sign language are naturally diverse; significant variability is introduced by elements like camera quality, variations in framing, and hand placement. Raw landmark coordinates were obtained using MediaPipe and subsequently passed through a series of normalization steps to make these differences smaller. Each set of coordinates was rescaled based on the frame size to make up for differences caused by the distance or angle of the camera. After normalization, temporal smoothing was used to get rid of the high-frequency jitter that is often seen in landmark detection outputs. This kept the natural flow of gestures without changing the motion patterns that underlie them.

Another problem that was fixed during preprocessing was the relatively small amount of annotated training data. To help generalize to new signers and environments, a number of data augmentation methods were shown. These included small changes in rotation to mimic natural changes in hand orientation, small changes in scaling to account for differences in hand size or camera angle, and horizontal mirroring to account for differences between left- and right-handed people. By adding these controlled perturbations during training, the network learns to prioritize invariant features. This makes it more robust in real-world situations where recording conditions are rarely the same.

After normalization and augmentation were done, each gesture sequence was turned into a set of 21 hand landmarks for each frame. This representation keeps the geometric relationships needed for accurate recognition while greatly reducing the number of dimensions compared to raw pixel data. It captures both fine-grained finger movements and more general hand positions. The compact feature format, which helps with computational efficiency, is directly in line with the system's sustainability goals. Classification accuracy is very important for public services and assistive technologies. Reduced memory use and lower inference latency make it possible for these technologies to work well on standard consumer hardware.

4.3 Model Configuration and Training Strategy

The sign recognition model in this study was developed using the Create ML framework and is predicated on ResNet50. A transfer learning approach was employed in conjunction with pretrained weights from the ImageNet dataset, which capture fundamental visual features such as edges and textures. This initialization set up a strong base that let the fully connected classification layer adapt to the unique spatial and temporal patterns of the chosen Word-Level American Sign Language (WLASL) gestures and fine-tune deeper residual blocks.

Iterative hyperparameter tuning was used to find the best possible balance between training stability and accuracy. A batch size of 32 produced effective gradient updates without consuming too much memory, according to empirical testing. The Adam optimizer, a popular algorithm for transfer learning tasks because of its adaptive step-size adjustments, was used to optimize the learning rate, which was set at 0.001. To avoid overfitting, training was carried out for a maximum of 50 epochs, with early stopping initiated when validation loss plateaued. For this type of multi-class classification problem, categorical cross-entropy was chosen as the loss function.

Every experiment was carried out on a typical workstation without GPU acceleration that had an Intel Core i7 processor (3.2 GHz) and 16 GB of RAM. This hardware selection was intentional because one of the study's objectives was to show that competitive recognition performance could be attained using widely available, low-end computer resources. Reaching this benchmark demonstrates how useful the framework is for implementation in real-world settings, especially in public service and educational settings where expensive hardware might not be accessible.

4.4 Evaluation Metrics

Four metrics commonly applied in multi-class classification tasks—recall, precision, accuracy, and F1 score—were employed to evaluate the effectiveness of the proposed framework. Accuracy indicated the proportion of samples assigned to the correct category, thereby reflecting the overall correctness of classification. Precision was calculated to determine the proportion of true positive predictions, thereby assessing potential bias toward false positives. Recall was measured to evaluate the system's ability to identify true positives and minimize false negatives. Lastly, the F1 score, which is the harmonic mean of precision and recall, showed how well the model worked in every way. This was especially helpful when the class distributions were split up but still had small differences.

In addition to these overall measures, a confusion matrix analysis was conducted to assess the performance of each class. This analysis was helpful because some signs, like "hello" and "good morning," have hand positions and paths that cross each other, which can be hard for people to see. The matrix put some gestures that were very similar to each other in the wrong group. This means that some things could be improved. The addition of facial expression cues or temporal attention mechanisms could facilitate the distinction between closely related signs.

The independent test set showed that the system was correct 94.0% of the time. The results showed that all gesture classes were recognized well and fairly, with precision, recall, and F1 scores of 92.0%, 94.5%, and 93.0%, respectively. For ten trials on the specified hardware setup, it took about 0.7 seconds per gesture to make an inference. Because the apps are responsive enough, schools and hospitals can use them. The system works even better in real life because it can always figure out what different types of signs and signer profiles mean.

Table 1 displays a summary of performance metrics from prior studies for comparative analysis with the current study. Although Aksoy et al. (2021) and Karakan et al. (2025) obtained marginally better accuracy on datasets that focused on alphabet gestures or used specialized hardware, these systems were either resource-intensive or lacked real-time capabilities. The framework presented in this study, on the other hand, is very suitable for deployment on consumer-grade devices in assistive and public service applications because it provides competitive accuracy while still being lightweight and energy-efficient.

Table 1. Comparative Evaluation of Sign Language Recognition Systems

Study	Dataset	Architecture	Accuracy (%)	Application Capability
Aksoy et al. (2021)	Custom dataset	CapsNet	99.7%	Limited
Oktekin et al. (2019)	Custom dataset	HMM	82%	Limited
This Study	Kaggle SL	ResNet50 + Landmark-Based	94.0%	Limited – But Integrable into Real-Time Systems
Karakan et al. (2025)	Custom dataset	YoloV8	90.7	Fully Real-Time

Additional information about the system’s functionality can be obtained by closely examining the confusion matrix shown in Figure 2. Generally speaking, most gesture categories were reliably identified; however, the misclassifications that did occur were not random but rather followed a clear pattern. In particular, mistakes tended to group together among gestures with similar motion paths or spatial arrangements. This observation suggests that these difficult cases may be the focus of future improvements. Integrating transformer-based temporal architectures or multimodal fusion techniques, which can model sentence-level dependencies and potentially lessen confusion between visually similar signs, could be one promising approach.

It also examined how data augmentation methods affected the overall performance of the system. Adding augmented samples made the model much better at adapting to changes in the position of the signer and the camera. This enhancement was less perceptible for gestures that are inherently ambiguous, as even augmented data could not fully resolve classification challenges. These results suggest a clear direction for future development: augmenting the dataset with more contextually rich samples and a broader array of sequential patterns could enhance the model’s capacity to manage subtle variations in sign articulation.

In conclusion, the proposed framework successfully achieves a balance between computational efficiency and recognition accuracy. This balance fits with the main goals of the study, which are to make communication technologies that are scalable and open to everyone, as well as to use eco-friendly computing methods.

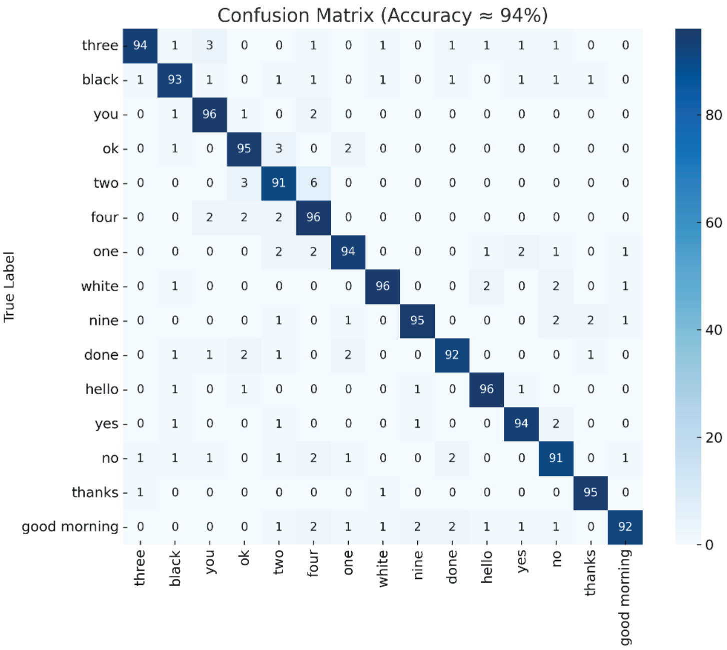

Figure 2. Applied Models' Confusion Matrix

5. DISCUSSION

The experimental results of the study demonstrate that even hardware with constrained processing capabilities can effectively recognise sign language. Older systems often needed strict lab conditions or special equipment to get an accuracy rate of about 94% without using special sensors. This is a big step forward. How quickly the model can make inferences is also very important. It takes an average of 0.7 seconds per gesture, which makes it easy for people to talk to each other in real life, like in classrooms and during clinical consultations, where delays can really slow down communication.

These results are good, but you should remember that there are some limits. The current system can only see one gesture at a time. It does not yet take into account that sign languages are always changing and have grammar that is hard to understand. It is important to use non-manual features like gaze direction, head tilts, and facial expressions to convey semantic meaning, but they are not modelled. This is what the confusion matrix shows.   Saying "hello" or "good morning" is a way to move through space, and people often get it wrong.

The technology in the development environment made this work go both faster and slower at the same time. Because Create ML was closed source, it was hard to try out new hyperparameters or change architectures. However, it made it easy to quickly prototype and deploy within the Apple ecosystem. Moving to open-source frameworks like PyTorch or TensorFlow would help the community make progress by making it easier to scale across platforms and allowing for more experimentation.

Another good thing about this framework is that it works well with sustainability goals. It is especially useful in places where resources are limited, like community health centers and rural schools, because it doesn't use much energy and works with consumer-grade devices, which saves money and protects the environment. This design philosophy is in line with the UN Sustainable Development Goals, especially those that focus on making technology and education available to everyone. It also fits in with current talks about AI that is good for the environment (Wright et al., 2023; Wu et al., 2021).

There are a number of obvious avenues for progress in the future. By incorporating sequence-aware models, like transformer-based attention mechanisms, the system may be able to process continuous signing and better capture linguistic context. To further enhance generalization to real-world situations, the dataset should be expanded to include a larger vocabulary, more diverse signers, and context-rich recording scenarios. Furthermore, adding multimodal cues, like upper-body posture and facial landmarks, would improve semantic representation and lessen misunderstandings between visually similar gestures.

Overall, this framework shows that without specialized hardware, accurate and energy-efficient sign language recognition is possible. With further improvements in sequential modeling, multimodal integration, and dataset expansion, this prototype could become a full-featured translation tool, greatly increasing accessibility for the deaf and hard-of-hearing community.

6. CONCLUSION

The point of this study was to see if regular video input and consumer-grade hardware could figure out sign language on their own. Using skeletal landmarks and a residual deep learning model, the framework could process each gesture in less than a second and get about 94% accuracy on the WLASL dataset. These numbers may not be the best in the field, but they are interesting because they came from real life, not a controlled lab.

The main focus of this work is on sustainability and inclusivity, which are very important. The system is flexible because it can work with a lot of different devices and does not need any special sensors or powerful GPUs. This framework is great for communities with few resources because it saves money and energy. The design choices made here are in line with global goals like the United Nations' Sustainable Development Goals, which stress how important it is for everyone to have equal access to technology and education.

There are still some big problems that need to be fixed. The current model can only understand separate gestures, not the continuous signing and grammar that real sign languages use. It is important to use facial expressions and head movements that do not involve hands to show meaning, but they should not be mixed together. To get around these limits, we need to use transformer-based models and multimodal data sources that are aware of sequences to improve semantic representation.

It is also getting harder to make datasets. If you add more signers, make the vocabulary bigger, and add scenarios that are full of context, the system will be more general and strong. If you switch from Create ML to open-source frameworks like PyTorch or TensorFlow, it will be easier to change the architecture, make the platform more scalable, and get the same results again.

In short, the study demonstrates that core hardware can support sign language recognition in a manner that is accurate, practical, and suitable for public use. The framework could turn into a full translation system that helps people who are deaf or hard of hearing talk to each other and get along better. This will depend on how well sequential modelling, multimodal integration, and expanding datasets get better in the future.

REFERENCES

Alhijaj, Jenan A. and Khudeyer, Raidah S. (2023). Techniques and applications for deep learning: A review. Journal of Al-Qadisiyah for Computer Science and Mathematics, 15(2).

Aksoy, B., Salman, O. K. M. and 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

Apple Inc. (2024). Create ML. https://developer.apple.com/machine-learning/create-ml/ adresinden 30 Mayıs 2025 tarihinde erişildi.

Das, T., Nayak, D. S. K., Kar, A., Jena, L. and Swarnkar, T. (2024). ResNet-50: The deep networks for automated breast cancer classification using MR images. 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 1-6. https://doi.org/10.1109/ASSIC60049.2024.10507980

Güney, S. and Erkuş, M. A. (2022). A real-time approach to recognition of Turkish sign language by using convolutional neural networks. Neural Computing and Applications, 34, 4069–4079.

Hasan, M. M. and Mishra, P. K. (2012). Hand gesture modeling and recognition using geometric features: A review of literature. International Journal of Computer Applications, 42(18), 1–7.

Kaggle (2025). Sign language recognition competition dataset. https://www.kaggle.com/competitions/sign-language-recognition adresinden 30 Mayıs 2025 tarihinde erişildi.

Karakan, A. ve Oğuz, Y. (2025). Real-time detection of Turkish sign language letters and numbers with deep learning. APJESS, 13(2), 31–41. https://doi.org/10.21541/apjess.1495405

El-Alfy, M. and Luqman, H. (2022). A comprehensive survey and taxonomy of sign language research. Engineering Applications of Artificial Intelligence, 114, 105198. https://doi.org/10.1016/j.engappai.2022.105198

Renjith, S., Rashmi, M. and Suresh, S. (2024). Sign language recognition by using spatio-temporal features. Procedia Computer Science, 233, 353-362.

Russell, S. J. and Norvig, P. (2020). Artificial intelligence: A modern approach (4th Ed.). Pearson Education.

Sutjiadi, R. (2023). Android-based application for real-time Indonesian sign language recognition using convolutional neural network. TEM Journal. pp. 1541-1549.

Oktekin, B. and Çavuş, N. (2019). İşitme ve konuşma engelli bireyler için işaret tanıma sistemi geliştirme. Folklor/Edebiyat, 25(97), 575–590.

Singh, J., Goyal, S. B., Kaushal, R. K., Kumar, N. and Sehra, S. S. (2023). Applied data science and smart systems. CRC Press. https://doi.org/10.1201/9781003471059

Takyi, K., Gyening, R.-M. O. M., Gueuwou, S. M., Nyarko, M. S., Adade, R., Borkor, R. N., Boadu-Acheampong, S. I. and Tabari, L. (2025). AfriSign: African sign languages machine translation. Discover Artificial Intelligence, 5 (6). https://doi.org/10.1007/s44163-025-00227-7

Veale, T., Conway, A. and Collins, B. (1998). The challenges of cross-modal translation: English-to-sign-language translation in the Zardoz system. Machine Translation, 13, 81–106.

Walde, A. S. and Shiurkar, U.D. (2017). Sign language recognition systems: a review. International Journal of Recent Research, 4(4), 451–456.

Wright, Dustin, Igel, Christian, Samuel, Gabrielle and Selvan, Raghavendra (2023). Efficiency is not enough: A critical perspective of environmentally sustainable AI. arXiv preprint arXiv:2309.02065. https://arxiv.org/abs/2309.02065

Wu, Carole-Jean, Raghavendra, Ramya, Gupta, Udit, Acun, Bilge, Ardalani, Newsha, Maeng, Kiwan, Chang, Gloria, Behram, Fiona Aga, Huang, James, Bai, Charles, Gschwind, Michael, Gupta, Anurag, Ott, Myle, Melnikov, Anastasia, Candido, Salvatore, Brooks, David, Chauhan, Geeta, Lee, Benjamin, Lee, Hsien-Hsin S., Akyildiz, Bugra, Balandat, Maximilian, Spisak, Joe, Jain, Ravi, Rabbat, Mike and Hazelwood, Kim (2021). Sustainable AI: Environmental implications, challenges and opportunities. arXiv preprint arXiv:2111.00364. https://arxiv.org/abs/2111.00364