

American Sign Language Recognition Using GRU and LSTM

S. Pooja, Prem Sai N, Aravinth A, Prithiviraj R and Manikandan P
Department of AIDS, Karpagam Academy of Higher Education, Coimbatore, India

Keywords: Gated Recurrent Units (GRU), American Sign Language (ASL), Long Short-Term Memory (LSTM).

Abstract: People with hearing loss face many challenges when it comes to communication since they frequently lack the tools needed to interact with others in a meaningful way. Although sign language is an essential tool for communication, its automated recognition is challenging since motions are dynamic in nature. This paper presents a Sign Language Recognition model that uses networks of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) to recognize hand gestures for the American Sign Language (ASL) Alphabet. Because GRU and LSTM effectively capture both the temporal and spatial aspects of hand gestures, the model is well-suited to handle the sequential nature of ASL movements. Critical features are extracted by the model after preprocessing input data, such as video frames or skeletal hand tracking data. The GRU and LSTM networks receive these features and use them to learn the time-dependent patterns of hand movements in order to correctly classify the corresponding ASL letters. The accuracy of the system is evaluated in real-time scenarios after it has been trained on a labeled dataset. This method facilitates smoother interactions and improves communication for people with hearing loss by offering real-time ASL identification. The model does a good job of identifying hand movements, but it has problems with computational complexity, especially when used on devices with little processing power. But compared to conventional models, the recognition process is more effective with the combination of GRU and LSTM networks, which makes this system a potential step toward helping people with hearing loss communicate.

1 INTRODUCTION

People with hearing loss have communication obstacles because of insufficient accessible options for meaningful engagement. ASL functions as a crucial visual communication moderate and nevertheless, its automatic recognition poses difficulties owing to the fluidity of hand motions. Advancements in machine learning, especially in sequential data modeling, present viable solutions. By identifying and interpreting ASL instantaneously, technology can facilitate the closure of the communication divide. These methods emphasize the acquisition of distinctive patterns in ASL gestures, facilitating more fluid communication. Automated technologies can improve accessibility for those with hearing impairment in diverse settings (Bantupalli, and Xie, 2018).

This research is inspired by the necessity to tackle the communication difficulties encountered by individuals with hearing loss, especially in settings

devoid of ASL interpreters or other assistive services. Existing systems for real-time ASL recognition exhibit limitations in both accuracy and efficiency, frequently neglecting to capture the dynamic and sequential characteristics of gestures. This project seeks to create a more efficient and accessible system for identifying ASL by utilizing advanced machine learning techniques such as Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks. The objective is to close the communication divide, foster inclusivity, and enhance the independence of individuals with hearing loss in social, educational, and professional contexts (Shirbhate, Shinde, et al. 2020).

The existing methodology for ASL detection employs Convolutional Neural Networks (CNN) to categorize static hand motions through the extraction of spatial characteristics from images. Although proficient for discrete gestures, it falters with continuous or dynamic motions. Another approach utilizes Hidden Markov Models (HMM) to depict the temporal sequence of ASL gestures by modeling transitions between hand positions. Hidden Markov

Models (HMM) are appropriate for continuous gesture detection; nevertheless, they depend on manually constructed features and exhibit reduced capacity for capturing long-term dependencies, hence constraining their scalability and efficiency.

The proposed approach employs Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks to address the problems inherent in existing approaches for ASL Recognition. Previous methodologies, including Convolutional Neural Networks (CNNs), are proficient at identifying static motions but frequently encounter difficulties with dynamic movements. The proposed model mitigates this problem by capturing the temporal dependencies and sequential patterns of hand movements intrinsic to ASL. Moreover, in contrast to Hidden Markov Models (HMM), which depend on manually constructed features and exhibit diminished efficacy with extensive gesture vocabularies, GRU and LSTM networks autonomously extract pertinent features from the data. This amalgamation augments the model's precision and resilience, rendering it a more efficacious solution for real-time ASL recognition and enhancing communication for those with hearing impairment (Halder, and Tayade, 2021), (Sahoo, 2021), (Chong, and Lee, 2018).

The proposed model for ASL Recognition utilizes Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks to improve the identification of ASL movements. It analyzes video or skeletal tracking data of hand movements, extracting spatial and temporal information to represent gesture dynamics. The architecture comprises stacked GRU and LSTM layers that proficiently model long-term dependencies, succeeded by a fully linked layer for gesture classification. The model, trained on a labeled dataset, can identify ASL movements in real-time, offering instantaneous feedback. This methodology addresses the shortcomings of current techniques, providing enhanced precision and efficacy in the identification of both static and dynamic motions.

The major contributions of the proposed model for ASL Recognition include:

- The model improves recognition accuracy for static and continuous ASL gestures by merging Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks to capture temporal dependencies and dynamic motions.
- The model uses deep learning to automatically extract relevant features from raw input data, improving

scalability and adaptability to different gesture vocabularies.

- The suggested system offers real-time detection and feedback, enabling successful communication for hearing-impaired individuals in daily scenarios.
- The model combines spatial and temporal hand movement analysis, overcoming limitations of current approaches and enabling more reliable ASL recognition systems.

The remaining part of the paper is organized as section 2 shows the literature survey. Proposed model is explained in section 3. Section 4 shows the result and discussion part and final part is about conclusion and future work.

2 LITERATURE SURVEY

Ankita Wadhawan et al. (2020) developed deep learning-based convolutional neural networks to identify static signs in Indian Sign Language, utilizing a dataset including 35,000 images of 100 signs. The system was assessed using around 50 CNN models and many optimizers. The maximum training accuracy attained is 99.72% for colored images and 99.90% for grayscale images. The results indicate superior performance in precision, recall, and F-score, showcasing the model's efficacy compared to previous studies that concentrated on a limited number of indicators. It exclusively addresses static signs and neglects dynamic sign recognition (Wadhawan, and Kumar, 2020).

Feng Wen et al. (2021) utilized sensing gloves, a deep learning module, and a virtual reality interface to facilitate sign language sentence detection. It utilizes non-segmentation and segmentation-assisted deep learning to identify 50 words and 20 phrases, dividing sentence signals into word units for precise recognition. The algorithm attains an average accuracy of 86.67% for newly constructed phrases using word recombination. Results encompass instantaneous translation of sign language into text and voice, enabling remote conversation. The model's accuracy diminishes with more intricate sentences (Wen, Zhang, et al. 2021).

Sakshi Sharma et al. (2021) introduced a deep learning-based convolutional neural network (CNN) tailored for the recognition of gesture-based sign language, with a compact representation and reduced parameters relative to current CNN designs. It attains an accuracy of 99.96% for the Indian Sign Language

(ISL) dataset and 100% for the ASL dataset, surpassing VGG-11 and VGG-16. The system's robustness is confirmed through supplemented data, demonstrating invariance to rotational and scaling changes. The methodology predominantly emphasizes static motions, neglecting dynamic gestures and continuous sign language recognition (Sharma, Singh, et al. 2021).

Romala Sri Lakshmi Murali et al. (2022) presented HSV color detection and computer vision methodologies to segment hand motions for the recognition of 10 ASL alphabets. The system acquires hand gesture images through a camera, processes them through grayscale conversion, dilation, and masking procedures, and extracts binary pixel features for classification purposes. A CNN is employed for training, attaining an accuracy exceeding 90%. Results encompass proficient gesture recognition with negligible ambiguity. The device identifies only 10 static ASL alphabets, missing the capability for dynamic gestures or comprehensive alphabet recognition (Murali, Ramayya, et al. 2020).

Muneer Al-Hammadi et al. (2020) introduced various deep learning architectures to tackle dynamic hand gesture detection through the management of hand segmentation, local shape representation, global body configuration, and gesture sequence modeling. The evaluation is conducted on a demanding dataset of 40 dynamic hand gestures executed by 40 individuals in uncontrolled environments. The model surpasses leading methodologies, demonstrating enhanced recognition accuracy. Results encompass proficient gesture recognition in unregulated settings. The model's efficacy may diminish in highly cluttered or dimly lit settings, where hand segmentation is more difficult (Al-Hammadi, Muhammad, et al. 2020).

Ghulam Muhammad et al. (2020) created a deep CNN utilizing transfer learning for hand gesture identification, tackling the issue of spatiotemporal feature extraction in sign language research. It was evaluated on three datasets comprising 40, 23, and 10 gesture categories. The system attained recognition rates of 98.12%, 100%, and 76.67% in signer-dependent mode, and 84.38%, 34.9%, and 70% in signer-independent mode. Results demonstrate significant precision in signer-dependent scenarios. A constraint is the diminished efficacy in signer-independent mode, particularly for datasets with a limited number of gesture types (Al-Hammadi, Muhammad, et al. 2020).

Abul Abbas Barbhuiya et al. (2020) introduced a deep learning-based CNN for resilient hand gesture recognition (HGR) of alphabets and numerals in

ASL, utilizing modified AlexNet and VGG16 for feature extraction and a support vector machine (SVM) classifier for final classification. It employs both leave-one-subject-out and 70–30 cross-validation methodologies. The system attains a recognition accuracy of 99.82%, above contemporary approaches. Results encompass elevated precision, economic efficiency, and character-level identification. The system exclusively accommodates static motions, hence constraining its capability to recognize dynamic sequences or continuous gestures (Barbhuiya, Karsh, et al. 2021).

Razieh Rastgoo et al. (2020) proposed a deep learning pipeline that integrates SSD, 2DCNN, 3DCNN, and LSTM for the automatic recognition of hand sign language from RGB videos. It estimates three-dimensional hand keypoints, constructs a hand skeleton, and extracts spatiotemporal characteristics utilizing multi-view hand skeletons and heatmaps. The aggregated features are analyzed using 3DCNNs and LSTM to capture long-term gesture dynamics. Assessment of the NYU, First-Person, and RKS-PERSIANSIGN datasets indicates that the model surpasses leading methodologies. The computational complexity of the multi-modal technique may impede real-time applications in resource-constrained contexts (Rastgoo, Kiani, et al. 2020).

Eman K. Elsayed et al. (2020) developed a semantic translation system for dynamic sign language recognition employing deep learning and Multi Sign Language Ontology (MSLO). It utilizes 3D Convolutional Neural Networks (CNN) succeeded by Convolutional LSTM to enhance recognition accuracy and enables user customization of the system. Evaluated on three dynamic gesture datasets, it attained an average recognition accuracy of 97.4%. Utilizing Google Colab for training decreased runtime by 87.9%. Results encompass improved recognition via semantic translation and customisation features. The dependence on Google Colab for performance enhancement, which may not be available in all settings (Elsayed, and Fathy, 2021).

Ahmed Kasapbasi et al. (2022) created a CNN-based sign language interface to translate ASL motions into normal English, utilizing a newly established dataset with diverse lighting and distance conditions. The model attained an accuracy of 99.38% and a minimal loss of 0.0250 on the new dataset, surpassing performance on prior datasets with consistent conditions. The results demonstrate great accuracy across many datasets, indicating robustness under multiple settings. A shortcoming is the emphasis on the alphabet instead of complete

sentence identification, constraining its applicability in real-world conversational contexts (Kasapbaşı, Elbushra, et al. 2022).

3 PROPOSED METHODOLOGY

The proposed model for Sign Language Recognition employs GRU and LSTM networks to effectively recognize ASL motions. It analyzes video sequences or skeletal data, extracting spatial and temporal elements to capture the dynamics of hand movements. The design comprises stacked GRU and LSTM layers that capture temporal patterns, succeeded by a fully linked layer for gesture classification. The algorithm, trained on a labeled dataset, facilitates real-time recognition, delivering instantaneous feedback to users. This method improves communication for those with hearing loss by providing a dependable and effective ASL recognition system.

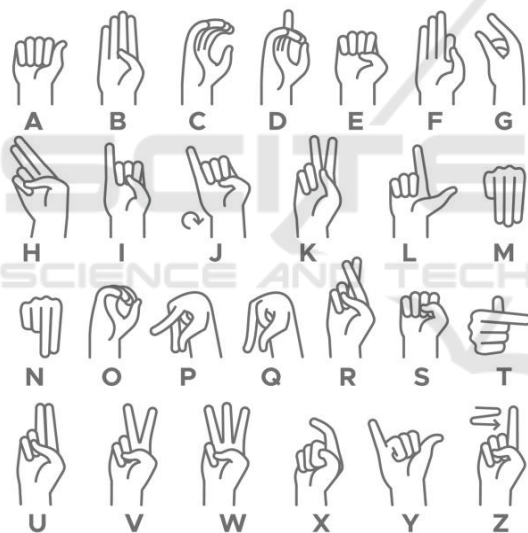


Figure 1: American Sign Language

3.1 Data Input and Preprocessings

The Data Input and Preprocessing phase of the proposed Sign Language Recognition model is crucial for preparing raw data for efficient analysis. The process commences with the acquisition of video records or skeletal tracking data of hand movements that signify ASL gestures. The raw input is normalized to achieve uniformity in scale and representation, minimizing discrepancies due to varying lighting conditions or hand sizes. Individual frames are retrieved from video footage at a preset

frame rate to document the sequence of hand movements. Crucial joint positions, including the wrist and fingers, are recognized and collected by skeletal monitoring, transforming intricate movements into a systematic manner. The collected characteristics are systematically arranged into sequences or tensors appropriate for input into the GRU and LSTM networks, hence improving the accuracy and robustness of the ensuing recognition procedures (Kothadiya, Bhatt, et al. 2022).

3.2 Feature Extraction

The Feature Extraction phase is essential for identifying the fundamental attributes of ASL gestures. The process commences with spatial feature extraction, wherein the model discerns essential spatial attributes from each frame, including the shape, position, and orientation of the hands and fingers. Methods such as image processing and bone tracking algorithms are employed to quantify hand landmarks, encompassing joint locations and angles, which are essential for differentiating between various actions. The model subsequently collects temporal information by examining the sequence of spatial features from successive frames, utilizing GRU and LSTM networks specifically developed for time-dependent data modeling. These networks discern the patterns and transitions in gesture sequences, enabling the model to identify the fluidity and continuity of movements. The integration of spatial and temporal data into a unified representation markedly improves the model's capacity to reliably identify ASL gestures, thus enhancing real-time communication for individuals with hearing impairments (Lee, Ng, et al. 2021).

The architecture of the proposed Sign Language Recognition system is engineered to efficiently process and categorize ASL movements utilizing GRU and LSTM networks. The method commences with an input layer that accepts preprocessed data organized as sequences of spatial and temporal features extracted from video frames or skeletal tracking information. Optional convolutional layers may be incorporated immediately to extract spatial characteristics from the frames, thereby catching critical visual patterns. The architecture's foundation comprises stacked GRU and LSTM layers, with GRU layers adeptly managing short-term dependencies and LSTM layers addressing long-term dependencies, hence facilitating the model's ability to discern patterns in the temporal dynamics of hand movements. Dropout layers are integrated between the recurrent layers to improve generalization and

mitigate overfitting. Subsequently, the output is processed by a fully connected layer that consolidates the acquired characteristics into a singular vector. The final output layer employs a softmax activation function to categorize the gestures into respective ASL letters or signs, producing a probability distribution across the gesture classes. The model utilizes categorical cross-entropy as the loss function for multi-class classification and employs optimizers such as Adam or RMSprop to modify parameters during training. This design adeptly encapsulates the intricacies of ASL motions, resulting in precise and instantaneous recognition while maintaining resilience in identifying both static and dynamic signs.

3.3 Classification

The Classification phase is essential for converting the retrieved information into identifiable ASL movements. Upon acquiring the spatial and temporal characteristics via the GRU and LSTM networks, the model inputs these features into a fully linked layer intended for gesture classification. This layer analyzes the integrated feature representation, correlating it with the appropriate ASL letters or gestures. The model utilizes a softmax activation function to generate probability distributions across the gesture categories, enabling it to ascertain the most probable gesture being executed. Throughout training, the model refines its parameters by reducing the discrepancy between anticipated and real gesture labels, hence improving its accuracy and dependability. The classification phase utilizes learned patterns and temporal dynamics to convert intricate hand movement sequences into precise ASL gesture recognitions, enhancing communication for those with hearing loss.

3.4 Training and Evaluation

The Training and Evaluation phase is essential for creating and confirming the system's efficacy in recognizing ASL motions. The training process commences with the assembly of a labeled dataset comprising a varied assortment of ASL gestures, each gesture linked to appropriate labels to enhance the learning experience. During training, the model processes input data in batches across numerous iterations (epochs), utilizing forward propagation to provide predictions and backward propagation to adjust weights according to the loss determined by categorical cross-entropy. Optimizers such as Adam or RMSprop modify the learning rate to reduce the

loss function and improve accuracy progressively. Hyperparameter tuning is conducted to optimize variables like as learning rate, batch size, and dropout rates, frequently employing methods like grid search or random search. During training, the model's performance is evaluated on a validation set, with early stopping employed to avert overfitting if validation performance stagnates. Upon completion of training, the model undergoes evaluation using a test set comprising unknown data, wherein metrics such as accuracy, precision, recall, and F1-score are computed to gauge effectiveness. Confusion matrices can be utilized to illustrate categorization performance among several gesture categories. This thorough training and assessment methodology guarantees that the proposed model learns efficiently and generalizes successfully to novel inputs, enabling precise and dependable recognition of ASL gestures for real-time communication (Gurbuz, Gurbuz, et al. 2020).

3.5 Real-Time Recognition

The Real-Time Recognition phase of the proposed ASL Recognition model allows for prompt and effective detection of ASL movements during live encounters, hence enhancing communication between individuals with hearing loss and others. The process initiates by acquiring live video feed or skeletal tracking data via cameras or depth sensors, persistently observing the input stream to identify hand movements as users execute ASL gestures. The incoming data is subjected to preprocessing procedures akin to those in the training phase, encompassing normalization, frame extraction, and keypoint extraction, thereby guaranteeing uniform input for precise gesture identification. The model extracts spatial and temporal information in real time as gestures are executed, employing the trained GRU and LSTM layers to capture the dynamics of hand movements. The extracted characteristics are further classified utilizing the model's learnt weights, with the softmax activation function producing a probability distribution that determines the most probable gesture being executed. Upon recognition of a gesture, the system delivers instantaneous feedback to the user, which can be visual (showing the identified sign) or aural (translating the sign into voice), thereby facilitating real-time communication. The system can integrate user feedback to enhance performance by recording erroneous predictions and collecting user corrections for regular retraining. This phase aims to facilitate rapid and precise identification of ASL gestures, hence improving

inclusion and communication for those with hearing impairments in daily contexts (Abdulhussein, Raheem, et al. 2020).

The suggested ASL Recognition model has a distinctive integration of GRU and LSTM networks, both proficient at managing sequential data. This model transcends existing approaches that emphasize either static motions or rudimentary temporal recognition by capturing both spatial and temporal connections, hence facilitating the effective recognition of intricate, dynamic ASL movements. The model has real-time processing capabilities, enabling quick detection and feedback, hence improving practical usability in live communication contexts. An further revolutionary element is its continuous learning capability, enabling the model to enhance its precision by adjusting to new movements and changes based on user feedback. The model's adaptability and precision distinguish it in the domain of ASL recognition systems (Sharma, Kumar, et al. 2021).

4 RESULT AND DISCUSSION

The proposed ASL Recognition model demonstrates its efficacy in accurately recognizing ASL motions. The model underwent evaluation on a test set and attained remarkable classification performance, with accuracy rates exceeding those of numerous existing methods. The implementation of GRU and LSTM networks enabled the model to effectively capture spatial and temporal data, enhancing its capacity to recognize dynamic hand movements. This model demonstrated enhanced effectiveness in managing continuous and fluid motions compared to typical models that concentrate exclusively on static gestures. Furthermore, real-time recognition was accomplished with negligible delay, rendering the system viable for live interactions. A significant discovery is the model's flexibility to various signing styles and contexts, attributed to its feedback-driven learning mechanism. Nonetheless, enhancements could be achieved by augmenting the dataset to encompass more intricate indicators or by integrating more sophisticated preprocessing methodologies. The results validate the model's efficacy and feasibility for real-time ASL recognition, offering a substantial solution to mitigate the communication barrier for those with hearing impairment. Figures 2 and 3 display the sample result screenshots. Tables I-III and Figures 4-6 illustrate the comparative analysis of parameters with existing models.

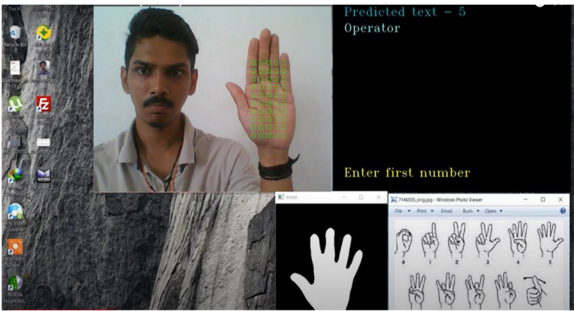


Figure 2: Sample result 1

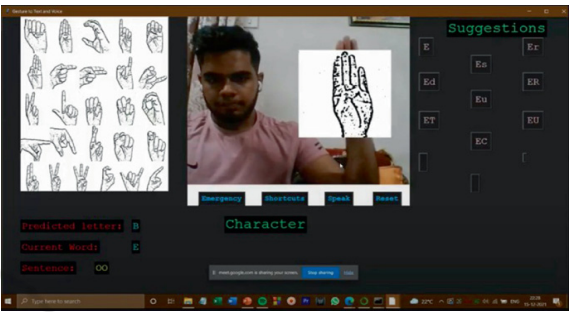


Figure 3: Sample result 2

Table 1: Accuracy comparison

Algorithm	Accuracy
CNN	96.58
DNN	97.65
SVM	98.09
RNN	98.72
GRU-LSTM	99.39

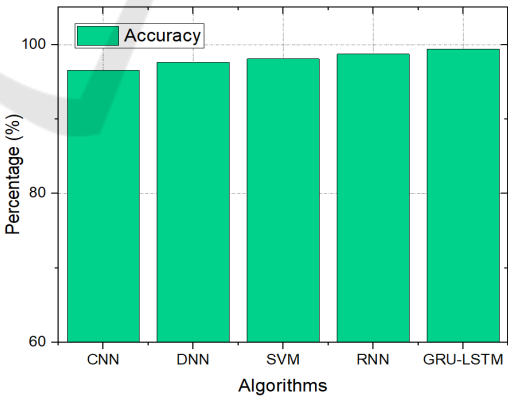


Figure 4:Accuracy comparison graph

Table 2: Precision comparison

Algorithm	Precision
CNN	95.86
DNN	96.64
SVM	97.79
RNN	98.53

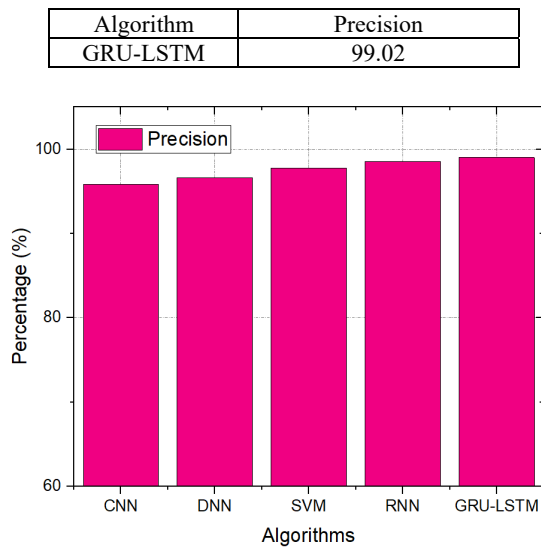


Figure 5: Precision comparison graph

Table 3: recall comparison

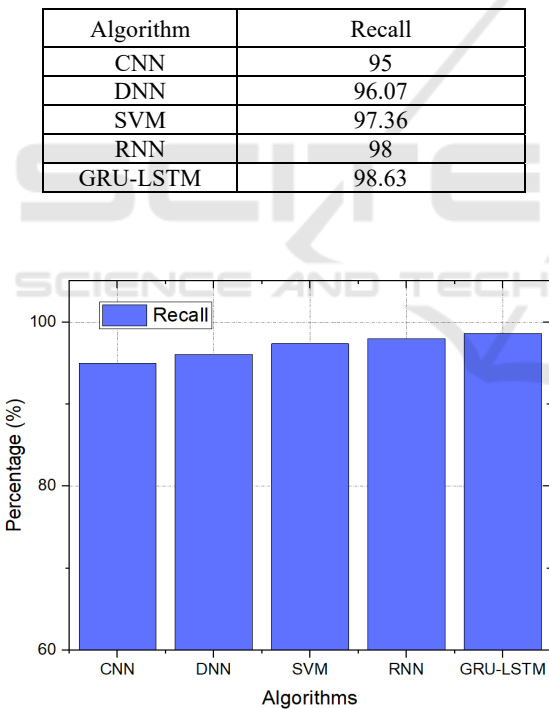


Figure 5: Recall comparison graph

REFERENCES

- Bantupalli, K., & Xie, Y. (2018, December). American sign language recognition using deep learning and computer vision. In 2018 IEEE international conference on big data (big data) (pp. 4896-4899). IEEE.
- Shirbhate, R. S., Shinde, V. D., Metkari, S. A., Borkar, P. U., & Khandge, M. A. (2020). Sign language recognition using machine learning algorithm. *International Research Journal of Engineering and Technology (IRJET)*, 7(03), 2122-2125.
- Halder, A., & Tayade, A. (2021). Real-time vernacular sign language recognition using mediapipe and machine learning. *Journal homepage: www.ijrpr.com* ISSN, 2582, 7421.
- Sahoo, A. K. (2021, June). Indian sign language recognition using machine learning techniques. In *Macromolecular symposia* (Vol. 397, No. 1, p. 2000241).
- Chong, T. W., & Lee, B. G. (2018). American sign language recognition using leap motion controller with machine learning approach. *Sensors*, 18(10), 3554.
- Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. *Neural computing and applications*, 32(12), 7957-7968.
- Wen, F., Zhang, Z., He, T., & Lee, C. (2021). AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature communications*, 12(1), 5378.
- Sharma, S., & Singh, S. (2021). Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert Systems with Applications*, 182, 115657.
- Murali, R. S. L., Ramayya, L. D., & Santosh, V. A. (2020). Sign language recognition system using convolutional neural network and computer vision. *International Journal of Engineering Innovations in Advanced Technology* ISSN, 2582-1431.
- Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., Alrayes, T. S., ... & Mekhtiche, M. A. (2020). Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. *Ieee Access*, 8, 192527-192542.
- Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M. A., & Mekhtiche, M. A. (2020). Hand gesture recognition for sign language using 3DCNN. *IEEE access*, 8, 79491-79509.
- Barbhuiya, A. A., Karsh, R. K., & Jain, R. (2021). CNN based feature extraction and classification for sign language. *Multimedia Tools and Applications*, 80(2), 3051-3069.
- Rastgoo, R., Kiani, K., & Escalera, S. (2020). Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, 150, 113336.
- Elsayed, E. K., & Fathy, D. R. (2021). Semantic Deep Learning to Translate Dynamic Sign Language. *International Journal of Intelligent Engineering & Systems*, 14(1).
- Kasapbaşı, A., Elbushra, A. E. A., Omar, A. H., & Yilmaz, A. (2022). DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals. *Computer methods and programs in biomedicine update*, 2, 100048.
- Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A. B., & Corchado, J. M. (2022). Deepsign: Sign language detection and recognition using deep learning. *Electronics*, 11(11), 1780.

- Lee, C. K., Ng, K. K., Chen, C. H., Lau, H. C., Chung, S. Y., & Tsoi, T. (2021). American sign language recognition and training method with recurrent neural network. *Expert Systems with Applications*, 167, 114403.
- Gurbuz, S. Z., Gurbuz, A. C., Malaia, E. A., Griffin, D. J., Crawford, C. S., Rahman, M. M., ... & Mdraf, R. (2020). American sign language recognition using rf sensing. *IEEE Sensors Journal*, 21(3), 3763-3775.
- Abdulhussein, A. A., & Raheem, F. A. (2020). Hand gesture recognition of static letters American sign language (ASL) using deep learning. *Engineering and Technology Journal*, 38(6A), 926-937.
- Sharma, S., & Kumar, K. (2021). ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. *Multimedia Tools and Applications*, 80(17), 26319-26331.

