Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 11 March 2026
Revised 20 April 2026
Accepted 18 May 2026

Natural Resources for Human Health 2026, Vol. 6 (1s): 585–600

KEYWORDS:
Hand gesture detection, MediaPipe, Sign language recognition, Text-to-speech, Visualization dashboard, Web-based assistive system

An Intelligent Web-Based Framework for Real-Time Indian Sign Language Recognition and Speech Translation with Interactive Learning and Performance Analytics

[1]S.D.Vijayakumar, [2]M.Prakash, [3]V.Gowri Shankar, [4]R.S.Kamalakannan, [5]Balasubramaniam C, [6]Karunakaran P, [7]D Nanthiya, [8]G.Brinda,

[1]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode, Tamilnadu-638052, India. mail2vijay.sd@gmail.com
[2]Department of Electronics and Communication Engineering, Kangeyam Institute of Technology, Nathakadaiyur, Kangeyam, Tirupur - 638108, India. prakashlakshmi333@gmail.com
[3]Department of EEE, Nandha College of Technology, Erode, Tamilnadu-638052, India. gowrishankar.v@nandhatech.org
[4]Department of ECE, Shree Venkateshwara Hi-Tech Engineering College, Gobichettipalayam, Tamilnadu – 638455, India. rskamal09@gmail.com
[5]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode, Tamilnadu-638052, India. cbalu.cse@gmail.com
[6]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode, Tamilnadu-638052, India. karuna.dr@gmail.com
[7]Department of CT-UG, Kongu Engineering College, Perundurai, Tamilnadu-638060, India. nanthiyaratha@gmail.com
[8]Department of ECE,M.P.Nachimuthu M.Jaganathan Engineering College,Chennimalai, Erode,Tamilnadu,India. bringandhi@gmail.com

ABSTRACT: The availability of accessible and interactive systems for sign language interpreting services between the deaf and hearing communities still poses a great challenge. To overcome this problem, in this paper, we are proposing a web-based framework which combines Indian Sign Language recognition, Text to Speech conversion, performance analytics and learning support in a single platform as SIGN2SPEECH. The proposed system uses the MediaPipe framework to extract the landmarks of the hands from both image, video, and real-time webcam feed, which allows for efficient gesture recognition with less complexity. For sign gesture classification, a light weight neural network is used and for the continuous recognition of sign gestures temporal filtering techniques are used to increase the stability of the prediction and decrease the number of errors in each frame. The developed framework allows multiple types of input such as upload images, upload videos, and interaction with the live webcam to communicate in sign language in real time. Through an integrated module that converts gestures into meaningful text and outputs into speech, they can be converted into meaningful text and corresponding speech outputs, thus to improve accessibility and communication effectiveness. The framework also includes a built-in analytics dashboard to provide visualization of model performance using confusion matrices, heatmaps, accuracy scores and other graphical tools, enhancing transparency and system evaluation.

Natural Resources for Human Health
eISSN: 2583-1194

Moreover, all the acknowledged results and confidence scores are accumulated into a structured JSON-based history module, that can be used for future reference and analysis. An interactive learning platform with A–Z alphabet sign and 117 frequently used words from a custom sign language dataset is also proposed to be incorporated as another contribution of this work. The experimental results show that the system has a high recognition rate with high real-time response and usability. SIGN2SPEECH is a scalable and intuitive solution as it integrates the functions of recognition, translation, visualization and learning into a web-based interface. The framework has great potential to improve communication, education and social inclusion of people who are hard of hearing.

1. Introduction

Sign Language Recognition (SLR) is the tool that can be used to achieve the effective communication between hearing-impaired people and the rest of the hearing population because it allows to decode the hand signs and motion patterns and to transform them into the linguistic tools that can be comprehended by the representatives of the hearing population. In spite of great improvements over the last few years, continuous SLR is still a difficult problem owing to the lack of segmentation of gesture streams, temporal correlations between signs, and lighting, background, and face variability. The first types of SLR systems were sensor based and handcrafted, non-invasive and hard to scale. Deep learning methods that operated on vision subsequently enhanced recognition rates, but tended to be computationally intensive so that real time operation on low-end systems was challenging. To overcome these shortcomings, hand keypoint-based landmark-based approaches have received interest, because they provide a dimensional reduction of the input, and the vital motion data is maintained. The extraction of hand landmarks using MediaPipe also facilitates the extraction of hand landmarks in a manner that is efficient and real time to be used in gesture recognition systems of light-weight applications. Nonetheless, the majority of current solutions are mostly only concerned with the single gesture recognition and do not offer much support to the stable continuous output, performance transparency, and user interaction. Inspired by these issues, this paper introduces SIGN2SPEECH, a web-based assistive system combining real-time sign language recognition and text and speech syntheses, visualization analysis and learning support to be implemented in practice.

The proposed system has an end-to-end workflow of sign language interpretation as illustrated in Figure. 1. It starts with the initial step where an image or video is captured either by uploaded material or by the camera techniques in real time. It goes through a feature extraction phase in which the appearance of pertinent features of the hand gestures will be obtained.

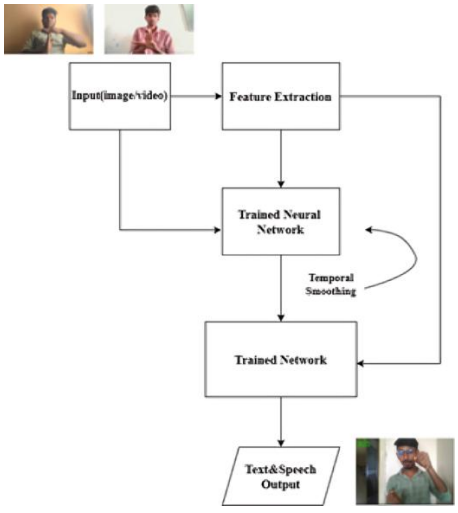

Figure 1: Training Procedure of the SIGN2SPEECH Model

They are then passed through a trained neural network to classify the gestures. Temporal smoothing is used to curtail frame level noise and misclassification to ensure that the recognition remains stable in continuous situations. The identified gestures are finally translated into meaningful text and speech outputs, which allows the users with hearing impairments to communicate with the rest of the population in real-time

2. Related Work

The development of sign language recognition (SLR) research has evolved from primitive sensor-based systems to vision-based machine learning approaches. Sensor-based methods, such as data gloves and motion sensors, were effective in controlled environments; however, they were intrusive, costly, and difficult to scale. Vision-based methods improved usability, but earlier handcrafted feature-based techniques were highly sensitive to lighting conditions, background variations, and signer differences (Priya & Sandesh, 2024; Das et al., 2024). The adoption of deep learning significantly enhanced SLR performance. Models such as convolutional neural networks (CNN), recurrent neural networks (RNN), and hybrid CNN-RNN architectures have proven effective in learning both spatial and temporal representations from image and video data (Singh et al., 2026; Kothadiya et al., 2024). More recently, transformer-based and attention-driven models have achieved high accuracy in continuous sign language recognition; however, they require large datasets and high computational resources, making them less suitable for real-time and resource-constrained environments (Khan et al., 2025; Nedungadi et al., 2025). Landmark-based recognition techniques using hand keypoints have gained attention due to their lower input dimensionality and robustness to environmental variations (Subramanian et al., 2022; Das et al., 2023). Frameworks based on MediaPipe enable efficient and real-time extraction of hand landmarks, making them suitable for lightweight SLR systems (Subramanian et al., 2022; Mariappan et al., 2024). However, most existing landmark-based approaches focus primarily on isolated gesture recognition and lack integration for performance analysis and user interaction (Yenisari & Yavuz, 2025). The proposed SIGN2SPEECH system addresses these limitations by integrating gesture recognition, text and speech generation, performance analytics, and learning support into a unified web-based platform.

2.1 Research Gap

- Most of the currently available methods are accurate and only work with isolated signs or alphabets and still lack the ability to understand continuous gestures and translate sentences.
- Some recent models can be based on advanced architectures like CNNs, Transformers, and multimodal approaches, although they require high computational complexity, which is hard to implement in a low-resource setting.
- Glove-based and sensor-based systems have been found to be more accurate when capturing gestures but have the disadvantage of added hardware requirements, making them more expensive and less convenient to users as well as less scalable.
- Though complex features can be dealt with by some vision-based techniques, the performance is usually influenced by environmental variations in terms of lighting, background noise and occlusions, and thus cannot maintain performance in real-life scenarios.
- Most of the currently available systems are primarily concerned with the accuracy of recognition, and they offer minimal end to end communication functionality as well as they are not well integrated to generate text, produce speech, and allow easy interaction with the user.

2.2. *Proposed Contributions*

- The main contributions of the proposed SIGN2SPEECH system are as follows: The proposed system supports real-time sign language recognition through image upload, video upload, and live webcam-based gesture analysis.
- The system utilizes MediaPipe-based hand landmark extraction for efficient and real-time gesture recognition with reduced computational complexity.
- A temporal filtering system is suggested to minimize the variability of the predictions and minimize the errors of continuous gestures recognition.
- The framework integrates text generation and text-to-speech (TTS) conversion to improve communication and accessibility for hearing-impaired individuals.

- It is made more practical, readable and usable through the creation of an interactive web interface that has various modules (upload, webcam, history, visualization and learning platform).
- This system is scaled and proved to be scaleable through the creation and use of a shared and recognizable set of 117 sign gestures.

3. Methodology

This section gives the methodology that has been used in the proposed SIGN2SPEECH system including the dataset preparation, preprocessing, model design, training strategy and post processing. The methodology is planned to produce the correct, steady, and real-time recognition of the sign language, and to be computationally efficient to implement it on the web.

3.1 Structure and Collection of Datasets

Figure. 2 displays sample images of custom sign language dataset designed to be used in SIGN2 SPEECH system. The gestures used are a smaller portion of the entire dataset to be illustrated. The entire dataset has more diverse sign gestures which were gathered by several participants.

Figure 2: Sample images from our own sign language dataset

The data was recorded in a different background condition, lighting conditions and hand orientation in order to increase model robustness and generalization. All the samples are manually marked and grouped into training, validation and testing sets. Besides gesture recognition, this self-acquired data has been utilized in the integrated sign language learning module of the proposed web platform too.

3.2 Image Enhancement and Preprocessing

In the proposed SIGN2SPEECH, the input images/video frames are first resized and converted into the RGB format in order to process the same number of input frames and images. MediaPipe framework recognizes hand parts and extracts hand landmarks which are 21 hand landmarks per frame. The shape of the hand and movement are modeled as a set of spatial coordinates that characterize these land marks. To be more resistant to scale and positional changes, the extracted landmark coordinates are scaled against a reference point on the hand. This kind of normalization is used in order to reduce the effect of camera distance variations and hand orientation variations. The resultant feature vectors are then presented to the gesture classification model which renders the recognition constant and efficient irrespective of the various conditions that exist in the real life.

3.3 Architecture of Models

SIGN2SPEECH system is a real-time sign language recognition system based on a lightweight architecture. The input images or videos are utilized in order to derive hand landmark features which give an effective representation of the hand gestures but maintain the necessary spatial information.

These landmark features are then presented to a well trained neural network in the gesture classification task as illustrated in Figure. 3 and the identified gestures then translated into a text and speech output. This practice minimizes the complexity of the computation and allows the efficient application in web based assistive applications.

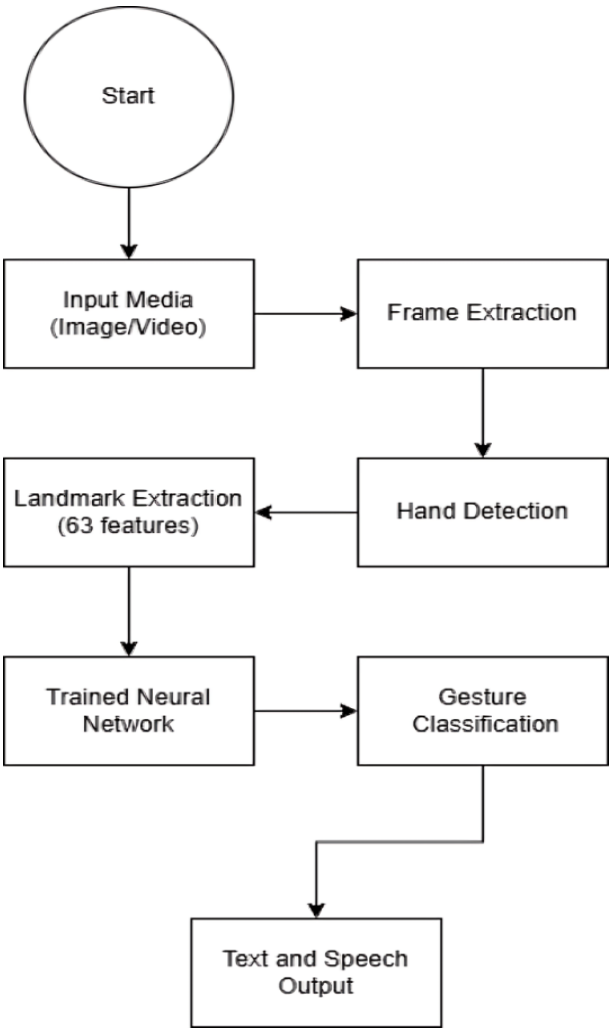

Figure 3: Architecture of the Proposed SIGN2SPEECH System

3.4 Sample Input Images

This section provides a sample input images that are applied in the proposed SIGN2SPEECH system. The samples contain various sign language gestures recorded on various users and are representative inputs to the recognition model.

The pictures exhibit a difference in the shape of hands, their orientation, and the background conditions which represents the conditions of real use. These sample inputs provide support in showing the heterogeneity of the dataset and confirming the strength of the system to identify the sign language signs in different visual conditions.

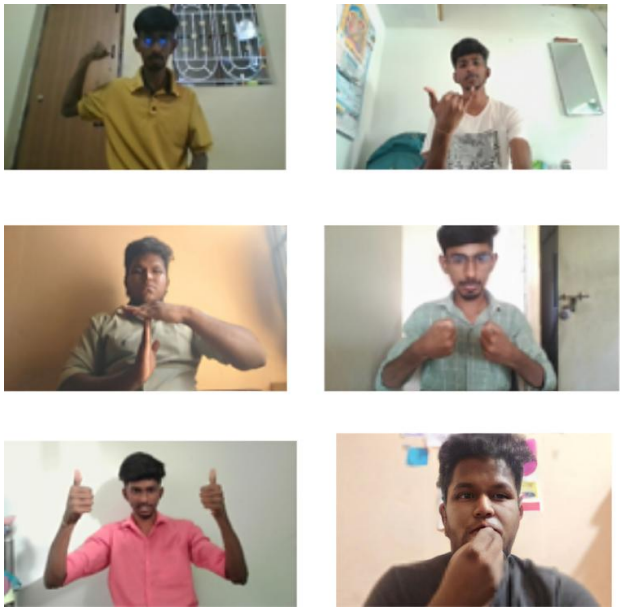

Figure 4: Sample Input Images of Sign Language Gestures

3.5 Method of Training

The SIGN2SPEECH model is trained using labeled samples of sign language gestures processed through MediaPipe, where hand landmarks are extracted. Each input frame is represented using 21 normalized hand landmarks, generating a compact feature vector that is used as input to the neural network classifier (Subramanian et al., 2022).

During training, the model learns gesture patterns by comparing predicted outputs with ground truth labels and updating its parameters to minimize classification error. To improve generalization and reduce overfitting, regularization techniques are applied. The system supports preprocessing of inputs from images, videos, and live webcam feeds, enabling real-time predictions (Chandan & More, 2025). Predictions are further refined using temporal aggregation, which smooths frame-level variations and improves consistency in continuous gesture recognition scenarios.

3.6 Temporal Filtering and Text-to-Speech Conversion

Frame-level predictions in continuous sign language recognition can vary due to hand movements and detection noise. The proposed SIGN2SPEECH system addresses this issue by applying temporal filtering, where predictions across consecutive frames are aggregated to produce a stable gesture output. This approach reduces abrupt misclassifications and ensures smoother and more reliable recognition results (Mariappan et al., 2024). Once a stable gesture label is obtained, it is translated into meaningful text, and a text-to-speech (TTS) module generates corresponding speech output. The integration of TTS enables visual signs to be converted into audible form, thereby improving accessibility and facilitating real-time communication for individuals with hearing impairments (Chandan & More, 2025; Najib, 2025).

The generated speech output has minimal latency, making the system suitable for real-time interaction. This combined text and speech generation mechanism enhances user comprehension and increases the applicability of the system in assistive and educational contexts (Yenisari & Yavuz, 2025).

4. Evaluation of the Proposed SIGN2SPEECH System

In this part, the proposed SIGN2SPEECH system is tested by standard classification measure and real time testing. The experiments are aimed at the accuracy of recognition, stability of the output and practical usability of the system in real life situations.

4.1 Performance of the Model

Table 1 shows the evaluation of the proposed SIGN2SPEECH system in the form of standard measures of classification. The system has an accuracy level of 92 percent and this means that most of the sign language gestures are classified correctly. It proves that hand landmark features used together with a small neural network are effective to identify gestures.

The accuracy rating of 92 improves the fact that the system does not make many false positive predictions, that is, gestures that are visually similar are not often classified as false. With a recall value of 91 percent, the model has been able to identify most of the actual sign gestures without leaving out any classes. The F1-score of 91 percent demonstrates equal efficacy of the model in terms of precision and recall, which proves that the model is very stable and consistent across the categories of gestures.

In general, Table 1 shows that the suggested SIGN2SPEECH system is able to recognize sign language fairly and precisely. The high results of all his assessment metrics indicate that the system is acceptable in real-time use and practical implementation in assistive and educational settings.

Table 1: Classification Performance for SIGN2SPEECH system

Metric	Value
Accuracy	92.00%
Precision	92.00%
Recall	91.00%
F1-Score	91.00%

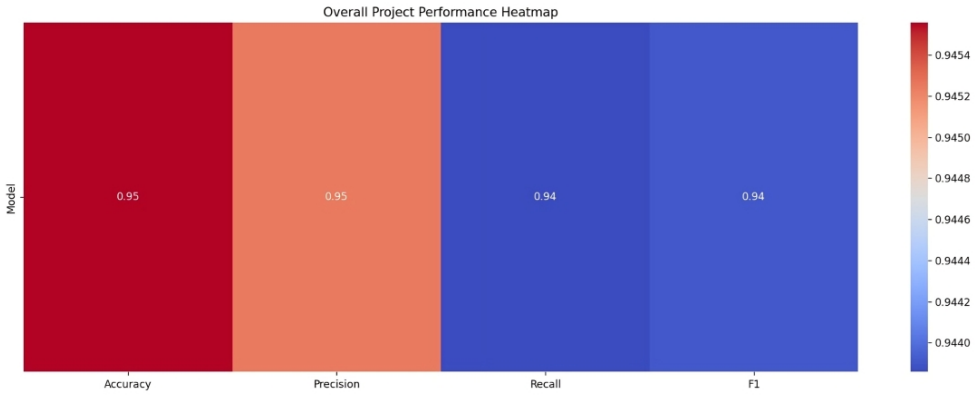

Figure 5: Overall Model Performance Heatmap

Figure. 5 shows the heatmap of overall performance of the proposed SIGN2SPEECH system, including such important evaluation indicators as accuracy, precision, recall, and F1-score. The high level of all these metrics at all times points to the fact that the system gives robust and efficient sign language recognition performance.

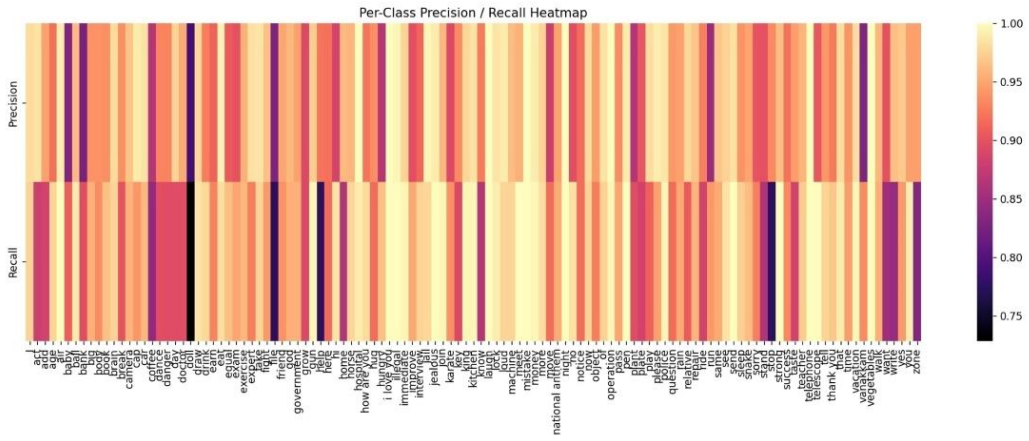

Figure 6: Per-Class Precision and Recall Heatmap

Figure. 6 presents the heatmap of per-class precision and recall of SIGN2 SPEECH system. The visualization has shown that there is a consistent performance among the majority of the gesture classes with some differences observed in few instances because some hand gestures are similar. On the whole, the findings show the strong stability and dependability of the model suggested.

4.2 Training and Validation Analysis

Figure. 7 illustrates the training and validation accuracy curve of the model proposed SIGN2SPEECH. The fact that the training and the accuracy of validation improves almost linearly with the number of epochs shows that the learning is successful and the performance in terms of generalization is good. The model attains a high value of accuracy that means that it is capable of learning meaningful gesture representations without overfitting. The condition between temporary variations is due to the disparities between the gesture patterns besides the real-time input circumstances.

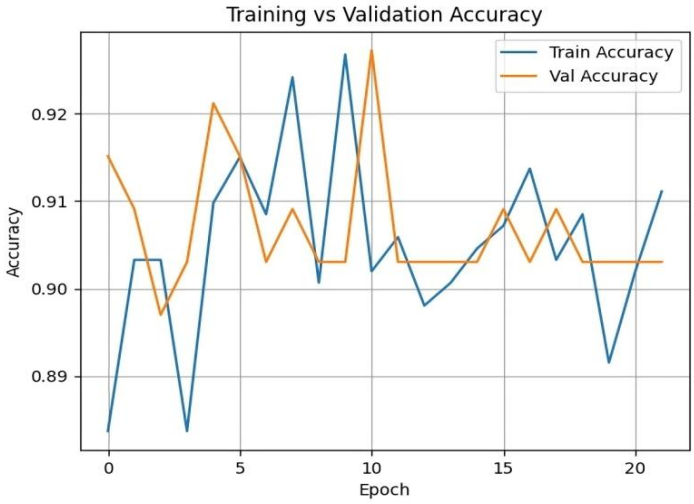

Figure 7: Training and Validation Accuracy Curve

Figure. 8 represents loss curve of training and validation of the proposed system. The reduced meeting of the two loss curves depict a consistent training behavior. The training and validation loss are comparable at the epochs and this fact demonstrates that the model can be applied to the unseen data. Such a trend of constant losses highlights the effectiveness of training strategy utilized in the SIGN2SPEECH system.

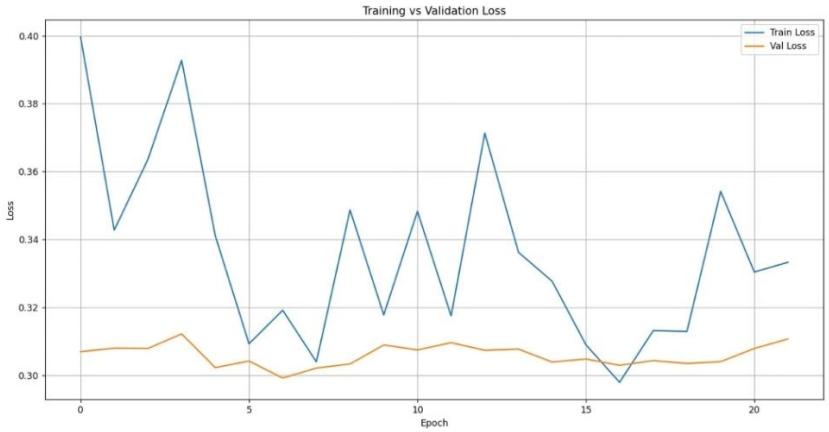

Figure 8: Training and validation loss curves

4.3 Gesture Recognition Accuracy in Real-Time Scenarios

Figure 9 illustrates the real-time performance of the proposed SIGN2SPEECH system using live webcam input. The system effectively recognizes hand gestures and corresponding signs in real-world scenarios (Mariappan et al., 2024). The results demonstrate consistent recognition of continuous hand movements. Landmark-based and temporal detection techniques help reduce sudden errors and improve prediction stability (Subramanian et al., 2022). These findings confirm that the proposed system performs reliably in real-time environments and is suitable for assistive communication applications.

Additionally, the system provides visual feedback with minimal delay, enhancing user interaction and overall usability.

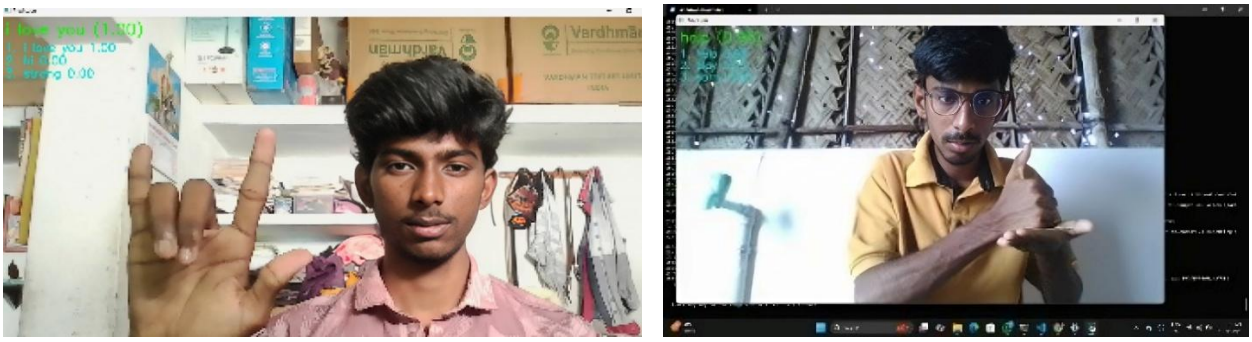

Figure 9: Real-Time Gesture Recognition Results

The system also performs well under varying lighting conditions and background environments, further demonstrating the robustness and feasibility of the proposed approach (Mariappan et al., 2024). This real-time performance highlights its effectiveness for everyday assistive and educational use.

4.4 Visual Output and Confidence Interpretation

The proposed SIGN2SPEECH system provides clear visualization of recognition results by displaying both the predicted gesture label and its corresponding confidence score. This graphical interface enables users to easily interpret system predictions in real time (Chandan & More, 2025).

The confidence score indicates the reliability of the recognized gesture and helps users assess the accuracy of the output. Higher confidence values represent greater model certainty, whereas lower values may result from environmental variations or similarities between gestures. By incorporating confidence-based visualization, the system enhances transparency, builds user trust, and improves overall usability in both assistive and educational applications (Yenisari & Yavuz, 2025).

4.5 Web-Based Practical Validation

The proposed SIGN2SPEECH system was also tried using its web-based application to examine its viability. The system was tried with different modes of input including uploads of pictures, recorded video input and live webcams. Figure 10(a) presents image-based gesture recognition, where uploaded images are processed and translated into corresponding text and speech outputs. Figure 10(b). illustrates gesture recognition of an input video that has been captured and it represents the fact that the system can capture continuous frames and provide continuous predictions. The input of image recognition is depicted in Figure 10(c). is a real-time gesture recognition with a web camera in which the system recognizes hand gestures and the system generates the output of the gesture in a high confidence. The system consistently produces reliable predictions with high confidence scores across different input modes. Interaction is also enhanced by the introduction of recognition, visualization and output modules to the web interface that makes it more accessible to the users. These results prove the notion that the proposed system could be successfully applied in the case of real- time conditions and be introduced into the assistive and educational programs.

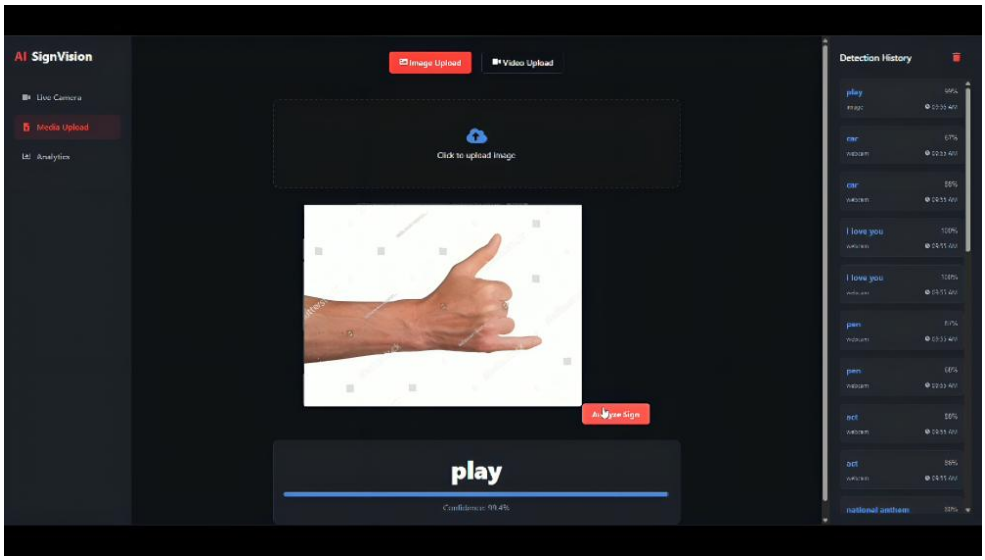

Figure 10(a): image-based gesture recognition

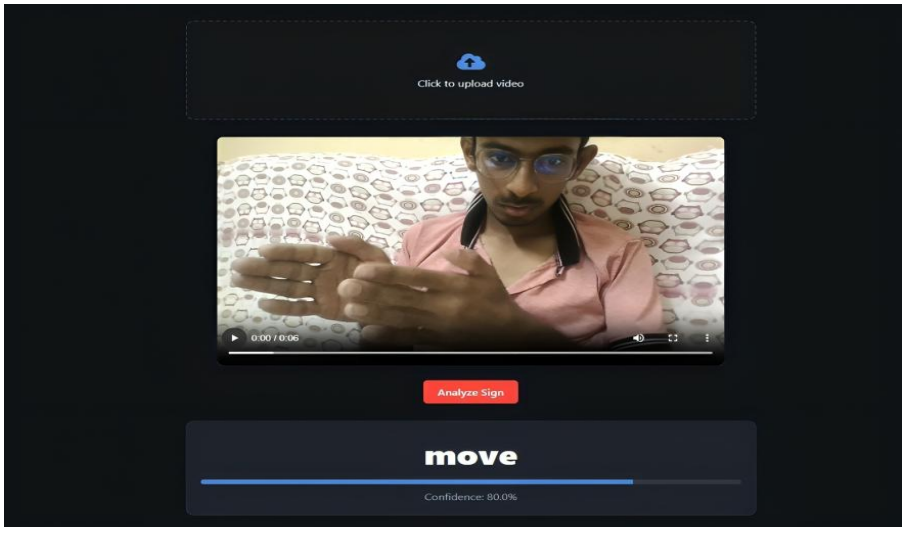

Figure 10(b): video-based gesture recognition

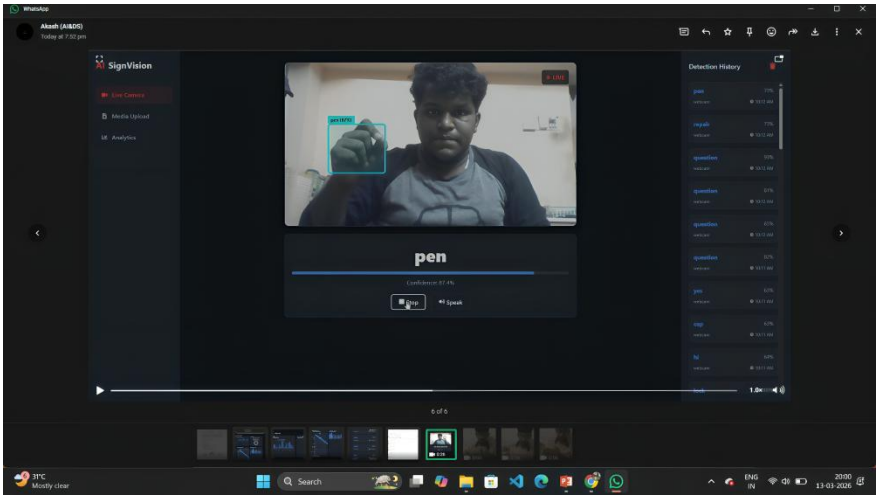

Figure 10(c): webcam input recognition

4.6 Limitations and Future Enhancements

Such aspects as high-speed hand movements, partial occlusions, and harsh light conditions may influence the performance of the proposed SIGN2speech system. Furthermore, the existing system offers a range of sign gestures, which is a limiting factor in terms of vocabulary. The ways to improve it in the future will be to add more vocabulary to the gesture and to enhance sentence-level comprehension to achieve more natural communication. Additional enhancements are also related to making the system stronger in the complicated environmental conditions and having a scalable system to be deployed on the web to make it more accessible and useful in real-life conditions.

5. COMPARISON WITH EXISTING SIGN LANGUAGE RECOGNITION APPROACHES

This section presents a comparative analysis of the proposed SIGN2SPEECH system with existing sign language recognition approaches. The comparison is based on key factors such as recognition accuracy, computational complexity, and real-time applicability.Recent attention-based deep learning models have demonstrated high accuracy; however, they often involve complex architectures and significant computational requirements (Yenisari & Yavuz, 2025). Similarly, glove-based systems provide real-time translation but depend on specialized hardware, limiting their scalability and practicality (Ram et al., 2025).

Hybrid approaches combining CNNs with large language models (LLMs) improve both recognition and language understanding, but they typically suffer from high computational cost and challenges in real-time deployment (Kumar et al., 2025). In addition, multimodal and transformer-based models enhance performance by capturing both spatial and temporal features; however, their high processing requirements make them less suitable for low-resource environments (Geetha et al., 2025; Singla et al., 2024).

In contrast, the proposed SIGN2SPEECH system adopts a lightweight, landmark-based approach using MediaPipe, enabling efficient real-time performance without the need for additional hardware. The system achieves a balance between accuracy and computational efficiency, making it well-suited for web-based and practical assistive applications (Subramanian et al., 2022; Das et al., 2023).

Table II: Comparison Table for Existing Proposed System

Method	Accuracy	Limitations
Attention-based DL Model	91%	High complexity, Requires sensors/gloves, limited scalability
Smart Glove System	89%	Hardware dependency, limited gestures
CNN + LLM Model	92%	High computational cost, limited real-time use
Multimodal / Transformer	90%	Complex architecture, high processing cost
Proposed SIGN2SPEECH	92%	Limited vocabulary (can be improved)

As Table II demonstrates, the current methods are highly accurate but usually use complicated models or extra equipment. Conversely, the suggested SIGN2SPEECH system is competitive in terms of accuracy, and low computation, and real-time operation. Such a balance renders it more applicable to practical and web-based applications of assistive sign language recognition.

6. RESULTS AND DISCUSSION

To validate the effectiveness of the proposed SIGN2SPEECH system for recognition and translation of Indian Sign Language, quantitative performance measure is used along with real time experimental analysis. The evaluation focused on measuring recognition accuracy, classification reliability, training behavior, real-time responsiveness, and practical usability. The proposed framework provides a reliable and efficient solution for sign language communication by integrating gesture recognition, text conversion, speech generation, and performance visualization within a unified web-based platform. The proposed system is able to recognize a wide range of Indian Sign Language gestures with an overall accuracy of 92%. The model achieved precision of 92%, recall of 91%, and F1 score of 91% in addition to the accuracy. The results of these values validate the effectiveness of the gesture classification model as well as its robust performance in preserving the balance between the different gesture categories. The effectiveness of the results obtained shows that the proposed framework is able to recognize hand gestures with a minimum number of classification errors, which is suitable for the practical use of the hand gesture for communication.

The high precision value assures that the number of false positive predictions is considerably lowered. A false positive in the sign language recognition system can cause misrecognition of sign language and miscommunication. Hence, the value of precision is an important factor to maintain reliability in the generated output. Likewise, the recall of 91% means the model is able to successfully capture the most of the real classes of gestures without neglecting important gestures. For assistive communication systems, it is especially significant that a high recall rate will not lead to the loss of messages or to the loss of information. The F1 score is another indicator of the strength of the model, as it gives a balanced measure of precision and recall. The similarity of these performance measures shows that the proposed system exhibits stable performance for various gesture sets and Inputs.

The training and validation analysis also proves that the learning strategy proposed in this research is effective. Both training and validation accuracy generally improve as the number of training epochs increases, indicating effective learning and good generalization capability. This trend shows that the model has learned meaningful representations of gestures from the hand landmark features it has extracted. Both training and validation curves are similar, as this indicates good generalisation ability and potential low risk of over fitting. Training overall performance consistency on both training and validation sets verifies the effectiveness of the architecture and training approach used. The curves for training and validation loss offer further insights into model stability. As training goes on, both loss values are continuously reduced and finally reach a minimum point, showing that the model is able to reduce the errors in classification. Loss curves are flat without significant variations indicating stable

learning behavior and efficient optimisation. The convergence properties suggest that the landmarks extracted by the MediaPipe models are informative and discriminative for gesture classification. Thus, reliable recognition performance can be obtained whilst maintaining computational efficiency from the proposed framework.

The evaluation process included several visualization techniques to get deeper insight on model performance. This heatmap of overall performance displays several metrics such as accuracy, precision, recall, and F1 score graphically. The consistently high scores measured in each of these measures reinforces the validity of the proposed framework. The visualization enables the user and the researcher to rapidly understand the performance of the model and to pinpoint strengths of the model's recognition process.

The per-class precision/recall heatmap provides a deeper view of the classification accuracy per gesture class. The results indicated that most of the gesture classes were recognized well and consistently. Some gestures, which have similar hand shapes or movement patterns have minor variations. This is because sign language recognition systems can vary due to the similarities between visually similar gestures, which can cause certain gestures to be misclassified. The results, however, are still quite satisfactory and show the strength of the landmark-based recognition strategy proposed.

One of the key aspects of the SIGN2SPEECH framework is its real-time capabilities. The system was tested with various input methods: live video stream from the webcam, uploaded images and recorded videos. Experimental results have shown that the framework can process inputs with very little latency with high recognition accuracy. The real-time webcam test showed that the system was able to detect, classify and translate hand gestures practically instantaneously to allow a smooth and natural communication. This feature is very important when using assistive devices for real-world tasks that demand real-time feedback.

Temporal filtering techniques are an important addition to the real-time performance. For continuous gesture recognition, the predictions at different frames can vary depending on the variation of hand movement, motion blur, or environmental noise at each frame. But if there is a need for temporal filtering, it can be done to combine the predictions in successive frames and give a more stable output. This means that accidental changes in the prediction are minimized and gesture recognition is more reliable. The findings of the experiments support the improvements in the consistency of the predictions and the user's experience throughout continuous communication sessions, provided by the temporal filtering.

The resistance of the proposed system was also tested at various environmental conditions. Different lighting conditions, background settings, hand orientations and user positions were used to test this framework. The results show that the MediaPipe landmark extraction technique does not show such a dependence on external changes as traditional image-based landmark extraction techniques. The system is mainly based on hand landmarks and not on the image pixels, which means it works well even under the varying environmental conditions. This feature could greatly enhance the relevance of the framework in real-world contexts.

One of the other key components in the proposed system is the integration of the text generation and speech generation modules. When the gesture is recognized, the text output is generated, and immediately converted into speech through the text-to-speech module. This conversion process is observed to have minimal delay, as indicated by the experimental observations, ensuring seamless communication between users. The speech output generated enables easy access to speech and helps hearing people hear the gestures of sign language without learning sign language. This will make the proposed framework more useful and enable inclusive communication.

The enhanced transparency and user trust is further improved thanks to the confidence score visualization module. The value of confidence is included for each prediction, indicating the level of confidence of the classification model. The higher the confidence value, the more reliable the prediction, and the lower the value means that there are similarities between the gestures or there are other factors that affect the prediction. Confidence information helps users to better understand the outputs of the system and provide an evaluation of the quality of the prediction. This transparency helps to make the system more usable and aids in human computer interaction. The integrated learning module also enhances the overall usefulness of the framework. The module also contains signs for alphabet (A-Z), and 117 frequently-used sign language words, allowing the learner to learn and practice sign language in the same application. The system integrates recognition, translation, visualization, and learning capabilities, making it more than just a recognition system; it is a learning and assistive system.

In general, the experimental outcomes and comprehensive performance evaluation show that the proposed SIGN2SPEECH framework is accurate, reliable, efficient and appropriate for practical application. This integration of gesture recognition, text-to-speech translation, performance analysis and learning support in a web based platform offers great benefits over traditional stand-alone systems The achieved performance metrics, stable training behavior, real-time responsiveness, and robustness under varying environmental conditions demonstrate the effectiveness of the proposed approach.

7. CONCLUSION AND FUTURE WORK

This paper introduced SIGN2SPEECH, a web-based Indian Sign Language recognition framework designed to reduce communication barriers between hearing-impaired individuals and the hearing community. The proposed system combines sign language recognition, Text Generation, Speech Synthesis, Visualization of performance and Learning Support in a single system. The framework efficiently detects gestures through high-quality hand landmark extraction using MediaPipe and a lightweight neural network classifier with low computation complexity. This allows for real time operation and running on the constrained devices.

The experimental evaluation results reveal that the proposed system performs well and provides high recognition rate with an overall accuracy of 92%, precision of 92%, recall of 91% and F1 score of 91%. The temporal filtering approach gives more stable predictions as it filters out the fluctuations at the frame level and minimizes misclassifications while performing continuous gesture recognition. Moreover, the built-in text-to-speech feature transforms recognized gestures into spoken text, enhancing accessibility and communication. The built-in visualization dashboard enables users to visualize the model's performance using heatmaps, confusion matrices, and evaluation metrics, making it easier to understand and transparent.

Another important feature of the framework is the alphabet signs A-Z and some frequently used words that are interactive and can be used to learn. This feature facilitates communication as well as sign language education, and is useful for learners and practitioners. The web-based architecture also enhances accessibility by making it easy for users to interact with the system, without needing specialized hardware. Although it works well, the current method has some drawbacks, such as a limited gesture vocabulary and problems when faced with very complex environmental conditions. The gesture set will be extended, real-time sign language sentence translation will be explored, and gesture recognition robustness will be enhanced in various scenarios. Other improvements like multi language speech generation, multi user support, advanced deep learning models, etc., can further enhance system performance. These developments will support the conversion of SIGN2SPEECH to a scalable, intelligent and practical assistive communication tool.

REFERENCE

[1] Bansal, N., & Jain, A. (2024). Word recognition from Indian Sign Language in videos using dual feature descriptor and GMT-Mask R-CNN. *Multimedia Tools and Applications*.

[2] Chandan, K. A., & More, V. B. (2025). Indian Sign Language interpretation using CNN and MediaPipe with text-to-speech integration. *International Journal of Cognitive Computing in Engineering, 6*(1), 55–65.

[3] Liu, L., Zheng, Y., Y., Zhang, X., & Yang, K. (2024). A Sign Language Recognition Based on Optimized Transformer Target Detection Model (pp. 197–208). https://doi.org/10.1007/978-3-031-50580-5_16

[4] Das, S., Biswas, S. K., & Purkayastha, B. (2023). Automated Indian Sign Language recognition system by fusing deep and handcrafted features. *Multimedia Tools and Applications*.

[5] Das, S., Biswas, S. K., & Purkayastha, B. (2024). Occlusion robust sign language recognition system for Indian Sign Language using CNN and pose features. *Multimedia Tools and Applications*.

[6] Dewangan, S., Patra, J. P., & Samal, S. (2025). Optimizing deep learning models for dynamic Indian Sign Language interpretation. *Social Science Research Network*. https://doi.org/10.2139/ssrn.5089115

[7] Geetha, M., et al. (2025). Toward real-time recognition of continuous Indian Sign Language: A multi-modal approach using RGB and pose. *IEEE Access, 13*, 60270–60283.

[8] Ghorai, A., et al. (2023). Indian Sign Language recognition system using network deconvolution and spatial transformer network. *Neural Computing and Applications*.

[9] Khan, A., Jin, S., Lee, G.-H., Arzu, G. E., Dang, L. M., Nguyen, T. N., Choi, W., & Moon, H. (2025). Deep learning approaches for continuous sign language recognition: A comprehensive review. *IEEE Access, 13*, 1–25. https://doi.org/10.1109/ACCESS.2025.3554046

[10] Kothadiya, D. R., Bhatt, C. M., Kharwa, H., & Albu, F. (2024). Hybrid InceptionNet based enhanced architecture for isolated sign language recognition. *IEEE Access, 12*, 90889–90902.

[11] Kumar, M., Natarajan, A., Visagan, S. S., Mahajan, T., & Sreeja, P. S. (2025). Enhanced sign language translation between American Sign Language and Indian Sign Language using LLMs. *IEEE Access*.

[12] Mariappan, S., Murugesan, P., & Selvan, H. M. (2024). Real-time interpreter for short sentences in Indian Sign Language using MediaPipe and deep learning. *Information Technology and Control, 53*(3), 888–898.

[13] Melanshia Violet, I. M., & Leena Sri, R. (2025). A comprehensive survey on recent advances and challenges in sign language recognition systems. *Discover Artificial Intelligence, 5*, 419.

[14] Najib, F. M. (2025). A multi-lingual sign language recognition system using machine learning. *Multimedia Tools and Applications, 84*, 27987–28011.

[15] Nandi, U., et al. (2023). Indian Sign Language alphabet recognition system using CNN with diffGrad optimizer and stochastic pooling. *Multimedia Tools and Applications*.

[16] Nedungadi, P., et al. (2025). Vision transformer-powered conversational agent for real-time Indian Sign Language e-governance accessibility. *Scientific Reports, 15*, 33055.

[17] Priya, K., & Sandesh, B. J. (2024). Developing an offline and real-time Indian Sign Language recognition system with machine learning and deep learning. *SN Computer Science*.

[18] Prudhvi, B., Neeraj, P., & Deepthi, V. H. (2023). An efficient real-time Indian Sign Language (ISL) detection using deep learning. *International Conference on Intelligent Computing and Control Systems*, 430–435. https://doi.org/10.1109/ICICCS56967.2023.10142596

[19] Ram, B. S., C. M., A. S., & Harikumar, M. E. (2025). Toward inclusive communication: Bit equivalent smart gloves for Tamil language finger spelling translation. *IEEE Sensors Letters*.

[20] Rastogi, U., Mahapatra, R. P., & Kumar, S. (2025). Advanced gesture recognition in Indian Sign Language using YOLOv10 with Swin Transformer. *Scientific Reports, 15*, 32894.

[21] Renjith, S., & Manazhy, R. (2024). Sign language: A systematic review on classification and recognition. *Multimedia Tools and Applications*.

[22] Shanmugham, M., Sekar, G., Kavitha, S., Vijayakumar, S. D., & Chithirai Pon Selvan, M. (2026). Integrating bioacoustic sensors for improved human–computer interaction. In K. Subbaraj, B. Balusamy, M. Chithirai Pon Selvan, F. Anish Mon, & C. Prabha (Eds.), *Computational bioacoustic artificial intelligence: Sustainable artificial intelligence-powered applications*. Springer. https://doi.org/10.1007/978-3-032-06088-4_11

[23] Sharma, S., & Singh, S. (2023). A spatio-temporal framework for dynamic Indian Sign Language recognition. *Wireless Personal Communications*.

[24] Singh, A., et al. (2026). A deep learning-based static gesture categorization and interpretation of Indian Sign Language using optimized GRU. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[25] Singla, V., Bawa, S., & Singh, J. (2024). Enhancing Indian Sign Language recognition through data augmentation and visual transformer. *Neural Computing and Applications*.

[26] Subramanian, B., Olimov, B., Naik, S. M., Kim, S., Park, K.-H., & Kim, J. (2022). An integrated MediaPipe-optimized GRU model for Indian Sign Language recognition. *Scientific Reports, 12*, 11964. https://doi.org/10.1038/s41598-022-15998-7

[27] Tao, T., Zhao, Y., Liu, T., & Zhu, J. (2024). Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges. *IEEE Access, 12*, 75034–75078.

[28] Wang, Z., Li, D., Jiang, R., & Okumura, M. (2025). Continuous sign language recognition with multi-scale spatial-temporal feature enhancement. *IEEE Access, 13*.

[29] Yenisari, E., & Yavuz, S. (2025). Deep learning-based sign language recognition using efficient multi-feature attention mechanism. *IEEE Access*.

[30] Deshpande, S., & Shettar, R. (2023). Hand Gesture Recognition Using MediaPipe and CNN for Indian Sign Language and Conversion to Speech Format for Indian Regional Languages. 1–7. https://doi.org/10.1109/csitss60515.2023.10334218