ISSN: 2182-2069 / E-ISSN: 2182-2077

Integrating CNN-LSTM Framework based Spatial-Temporal Analysis for Effective Sign Language Gesture Recognition in Forensic and Security Domains

Dr.G. Geetha[1*], Dr.J. Godwin Ponsam[2], Dr.V. Elizabeth Jesi[3], Dr.M. Mahalakshmi[4], Dr.S. Thenmalar[5], and Dr.P. Mahalakshmi[6]

[1*]Assistant Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. geethag@srmsit.edu.in, https://orcid.org/0000-0002-1361-9123

[2]Associate Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. godwinj@srmist.edu.in, https://orcid.org/0000-0001-6695-0568

[3]Associate Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. jesiv@srmist.edu.in, https://orcid.org/0000-0001-7797-2586

[4]Assistant Professor, Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. mahalakm5@srmist.edu.in, https://orcid.org/0000-0002-5222-5549

[5]Associate Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. thenmals@srmist.edu.in, https://orcid.org/0000-0003-2724-8711

[6]Assistant Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India. mahalakp@srmist.edu.in, https://orcid.org/0000-0002-0179-5699

Received: October 20, 2025; Revised: November 26, 2025; Accepted: January 16, 2026; Published: February 27, 2026

Abstract

Gesture recognition based on sign language is significant in forensic investigations and security surveillance, especially when monitoring non-verbal communication in sensitive settings is required. Nevertheless, current recognition systems often struggle with complex spatiotemporal dependencies and real-world variations, such as lighting, occlusion, and signer variation. To overcome these issues, this paper will develop a combined Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) system that leverages spatial-temporal features of sign language to analyze sign language gestures for forensic and security applications. The proposed methodology uses a deep CNN to extract high-level spatial features of video frames, including hand shape, orientation, and movement patterns. An LSTM network is then used to process these spatial features in sequence to model temporal dynamics in a gesture sequence. The framework has been tested on benchmark sign language video datasets containing more than 25,000 gesture samples across various

Journal of Internet Services and Information Security (JISIS), volume: 16, number: 1 (February-2026), pp. 797-810. DOI: *10.58346/JISIS.2026.I1.046*

*Corresponding author: Assistant Professor, Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.

sign categories. The performance measures were accuracy, precision, recall, and F1-score. The results of the experiment show that the CNN-LSTM model achieved higher overall recognition rates (94.8) than the traditional CNN-only and handcrafted feature-based models, by 8.6 and 14.2, respectively. The proposed system was also found to be more robust under noisy, low-resolution conditions, achieving an accuracy of 92.1 in simulated surveillance. Moreover, misclassification of visually similar gestures was reduced by 31 % through temporal modeling with LSTM. Finally, the integrated CNN-LSTM is a robust, scalable system for recognizing sign language gestures in law enforcement and security applications. Its capability to capture spatial and temporal properties enhances its interpretability and operational effectiveness, thereby boosting its application in automated surveillance, evidence analysis, and inclusive security communication systems.

Keywords: Sign Language Gesture Recognition, CNN–LSTM Framework, Spatial–Temporal Analysis, Deep Learning, Forensic Applications, Security Surveillance Systems, Human–Computer Interaction.

1 Introduction

Sign language gesture recognition (SLGR) is the automatic interpretation of hand gestures, body language, and related visual representations into useful linguistic meaning. It is an essential interface between the digital/human communication systems and hearing-impaired people. In addition to assistive communication, SLGR has become increasingly significant in forensic investigation and security surveillance, where non-verbal communication can often convey intent, coordination, or hidden messages. In both controlled and uncontrolled conditions, the accuracy of dynamic gesture decoding in video streams improves situational recognition and promotes evidence-based analysis (Sarowar et al., 2025; Rathipriya & Maheswari, 2024). With the use of intelligent video analytics in the development of security systems, there has been a strong need for powerful gesture recognition models to explain complex human behavior in real time.

Even with significant improvement, SLGR remains technically daunting because gestures involve spatial-temporal complexities. Differences in hand shape, motion speed, occlusion, signer position, and environmental factors generate significant intra- and inter-class variability. Conventional vision-based systems employed manual features that are difficult to generalize across datasets and real-world scenarios (Luqman & ELALFY, 2022). Moreover, statistical frame analysis cannot capture the temporal dependencies that are important for separating visually similar gestures that differ primarily in motion path or sequence order. The issue is also complicated by the space of recognized constant signs, which affects boundary detection and co-articulation among gestures (Rastgoo et al., 2025). Recent reviews also note that improvements in spatial or temporal modeling alone are insufficient to achieve high-accuracy recognition in real-world conditions (Sarowar et al., 2025; Baihan et al., 2024).

Figure 1(a) shows a top-level view of the entire sign language gesture recognition pipeline, which sequentially transforms raw video input into a gesture label. It demonstrates how continuous sign language video streams are initially converted into individual frames. Then they are preprocessed, that is, they are resized, normalized, and noise is filtered to improve the visual quality. Feature-extraction modules then process the polished frames to extract discriminatory spatial features, which are later fed to a gesture classification model to determine the sign associated with them. The figure also provides a conceptual overview of the end-to-end process of automated sign language gesture recognition systems.

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have become a widely popular solution to these problems, in the form of hybrid deep learning architectures. CNNs are good at learning hierarchical spatial representations of hand shapes, textures, and the

orientation of frame images, whereas LSTMs are good at capturing long-range temporal correlations in sequential data. The complementary design enables end-to-end learning of space-time information without manual feature engineering (Aicha et al., 2024; Sukhavasi et al., 2024).

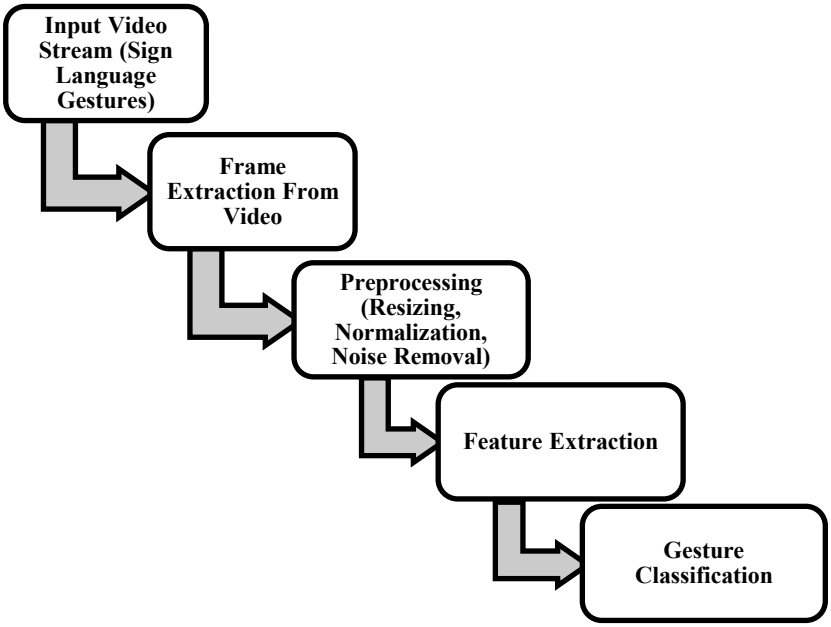

Figure 1(a): Overview of sign language gesture recognition pipeline

Empirical evidence shows that CNN-LSTM models are consistently superior to single CNN or RNN models, especially for dynamic gesture recognition in real-world settings (Baihan et al., 2024). Sensor fusion, better learning strategies, and extensions of cross-modal cues also improve robustness and scalability (Kanwal & Altaf, 2025; Fang et al., 2025). CNN-LSTM architectures are appropriate for analyzing complex temporal patterns, as their success is supported by related tasks such as deepfake detection (Khyati et al., 2025).

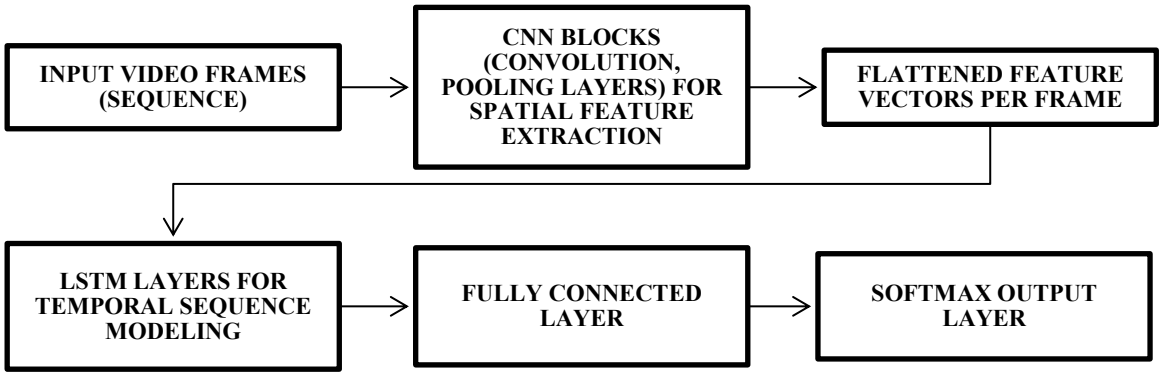

Figure 1(b): Video action recognition using CNN-LSTM architecture

The diagram (Figure 1(b)) shows a hybrid CNN-LSTM video action recognition system, with the input video frames first fed to CNN blocks (convolution and pooling layers) to obtain spatial features, which are then flattened into feature vectors. The vectors are fed into LSTM layers to capture temporal dependencies within the sequence of frames. Lastly, the processed features are fed through a fully connected layer and a softmax output layer to identify the action in the video.

The main issue this paper addresses is the lack of reliable, accurate spatial-temporal gesture recognition systems suitable for forensic and security settings. Finding a solution to this issue is crucial to the development of intelligent surveillance, the enhancement of non-verbal communication analysis, and the inclusion of security systems when gestures are essential to the interaction.

The paper introduces a technically refined CNN-LSTM-based spatial-temporal framework for dynamic sign language gesture recognition. The proposed approach will focus on robust feature mining, sound temporal modeling, and suitability for forensic and security applications, thereby overcoming the weaknesses of current hybrid architectures. By aligning real-world operational constraints with the architectural design, this work is a step forward toward the practical deployment of deep learning-based SLGR systems.

The rest of this paper is organized as follows. Section II provides a review of related work in sign language gesture recognition, focusing on the current spatial-temporal analysis methods and deep learning architectures based on hybrid methods. Section III outlines the proposed methodology, which comprises preparing the dataset, designing the CNN-LSTM architecture, and defining performance measures to evaluate the model. Section IV will cover the results of the experiment, in which the model is quantitatively and qualitatively analyzed in terms of performance and compared with baseline approaches. Lastly, Section V wraps up the paper by summarizing the main findings, emphasizing the suggested framework, and providing possible future research directions.

2 Literature Review

Early studies in sign language and gesture recognition focused primarily on vision-based machine learning methods that used handcrafted spatial features and shallow classifiers. Although these techniques provided computational ease, they failed miserably when subjected to real-world variations such as illumination, background clutter, and signer differences. New research has been directed towards a deep learning-based framework that allows for the automatic extraction of features and enhances generalization. Human activity surveys and gesture recognition point out that the replacement of traditional classifiers with deep neural networks is an essential step toward obtaining a greater recognition accuracy (Saleem et al., 2023; Khan, 2022). Alongside, other studies in the lip-reading and facial emotion recognition domains have also shown that visual-only modalities require great spatial modeling to deal with the subtle differences in motion as well as semantic ambiguity (Dixit et al., 2024; Khan, 2022). These results point to the need of learning complex visual representations in sign language interpretation through architectures.

One of the main areas in contemporary gesture recognition systems is the use of spatial-temporal modeling as a research topic. The CNN-based models have a rich usage in the extraction of spatial features, whereas the temporal dependencies are represented with the help of recurrent architectures or temporal convolution. Radar and wireless sensing methods also diversify the spatial-temporal analysis of vision to show that deep learning is versatile concerning the modeling of motion dynamics (Shen et al., 2022; Mosharaf et al., 2024). Fusion of multi-modal signals including visual, radar and WiFi signals demonstrated enhanced resilience in complex settings, especially when it comes to situations that involve occlusion (Wang et al., 2024). The ResMFuse-Net network, which requires the multi-level spatial-temporal fusion to provide fine-grained activity recognition, is an example of the efficiency of hierarchical feature aggregation (Asif et al., 2024). In addition, ARNet proposes joint space-temporal deep learning pipelines in action recognition, which supports the significance of co-dimensional feature education (Dawood et al., 2025).

CNN-LSTM architecture has become highly popular because of its capability to separate the spatial and temporal learning, keeping the end-to-end optimization. The CNN layers are effective in encoding the spatial representations of hand shapes and posture positions whereas the LSTM units maintain the temporal continuity and long-term connections. The design has been shown to be useful in different biometric and behavioral recognition problems, such as gait recognition and action recognition (Tiwari et al., 2025; Dawood et al., 2025). CNN-LSTM models provide a good trade-off between the performance and the computational requirements compared to purely convolutional or transformer-based models. However, limitations remain. LSTMs can handle quite long sequences, but they cannot tolerate noisy time sequences, and thus require attentive pre-processing and sequence-cutting. Moreover, recent surveys also suggest that further improvement in adaptation in changing environments can be achieved by hybrid systems combining soft computing or multi-agent learning (Houssein et al., 2025). In spite of these difficulties, CNN-LSTM still serves as an effective baseline to the modeling of gestures in space and time.

The literature review shows that to achieve a good sign language gesture recognition, a strong spatial-temporal modeling that can deal with dynamic motion, environmental variability and semantic complexity are necessary. Although other modalities of sensing and combination methods keep improving, CNN-LSTM systems have proven and established themselves as a reliable solution to sequential visual analysis. These results are also directly relevant to the proposed study as they support the use of CNN-LSTM frameworks and point to the areas of architectural optimization and application-specific optimization.

3 Methodology

Data Collection and Preprocessing for Sign Language Gesture Dataset

Data that was employed in this experiment is a videotape setup of dynamic sign language components used by several participants in the changing light and background conditions. Every series of gestures is captured at a constant frame rate in order to maintain the time scale. All the videos are resized to the standard spatial resolution and normalized to minimize the variance in illumination to make sure that there was consistency in samples. Frame-level preprocessing involves background suppression by means of adaptive thresholding and noise removal by the use of a Gaussian filter. A set of motion-based segmentation and bounding box extraction is used to localize hand regions in order to reduce extraneous spatial information. The temporal preprocessor uses continuous videos and divides them into a series of fixed length gestures. Given a gesture video as a sequence of video frames $V = (F_1, F_2, \ldots F_T)$, T is the number of video frame. The intensities of pixels of each frame Ft are normalized, i.e. pixel intensities are between [0,1]. Data augmentation methods that include horizontal flipping, temporal jittering, and minor rotation are utilized to enhance model generalization, as well as reduce cases of overfitting. The last dataset is subject-independently split into training, validation and testing subsets to perform an unbiased evaluation.

This scheme is a video-based analysis workflow (Figure 2), in which input datasets are processed to be in a preprocessing phase, including frame normalization and segmentation. A Convolutional Neural Network (CNN) is then used to extract features on each frame and LSTM processing is used to record sequential dependencies. The output feature sequences are inputted into a classification and prediction module, and a performance evaluation stage is then an evaluation of the accuracy and effectiveness of the model is evaluated.

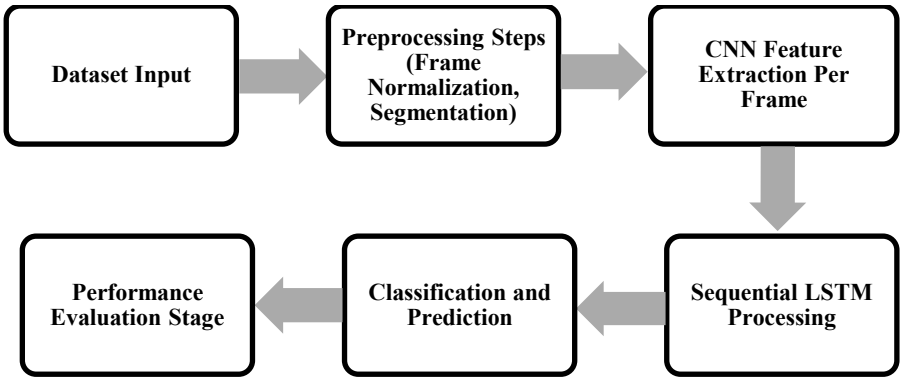

Figure 2: CNN-LSTM based video frame analysis pipeline

Implementation of CNN–LSTM Framework for Spatial–Temporal Analysis

In the proposed model the Convolutional Neural Network is used as a spatial feature extraction network and a Long Short-Term Memory network is used as a temporal modeling network. In every frame F_t, CNN is trained to obtain a hierarchical representation in terms of stacked layers of convolutions and pooling. The spatial feature of time t is represented as defined in Equation (1):

$$s_t = f_{CNN}(F_t) \quad (1)$$

$f_{CNN}(\cdot)$ is the convolutional mapping. The resultant spatial features are sequentially fed into LSTM in order to learn the temporal dependencies. LSTM computes its hidden state based on gated operations which are formulated as follows, shown in Equation (2):

$$h_t = f_{LSTM}(s_t, h_{t-1}) \quad (2)$$

Above, h_t denotes the hidden state at time t. The prediction of the final gesture is made by users with a softmax classifier using the final state of hiding, defined in Equation (3):

$$\hat{y} = softmax(Wh_T + b) \quad (3)$$

W and b are parameters that are learned. The end-to-end training of the model is based on categorical cross-entropy loss minimized with the help of the Adam algorithm. The techniques included are dropout and batch normalization that enhance the stability of convergence and removes overfitting.

Proposed CNN–LSTM Algorithm

Algorithm 1: CNN–LSTM Sign Language Gesture Recognition

Input: Video sequence V = {F1, F2, ..., FT}

Output: Predicted gesture label ŷ

1: Preprocess each frame Ft

2: for t = 1 to T do

3: st ← CNN(Ft)

4: ht ← LSTM (st, ht−1)

5: end for

6: ŷ ← Softmax (W hT + b)

7: return ŷ

This algorithm outlines the overall end-to-end process of sign language gesture recognition of video sequences based on the proposed CNN-LSTM model. The videos are then preprocessed to improve on the hand features that are of interest and reduce the level of noise. Gesture-specific visual patterns are then learned as the spatial features are extracted out of individual frames using the convolutional neural network. All these properties are gradually propagated through the network of LSTM that captures the temporal relationships between different frames to maintain motion continuity and gesture dynamics. Lastly, the obtained temporal representation is categorized by a softmax layer to generate the estimated gesture label and this allows successful categorization of dynamic sign language gestures.

Assessment Measures of Appraising the Success of the Framework

Various evaluation metrics are used in order to be able to measure the performance of the model quantitatively. Classification accuracy is used to determine how well the predictions are correct on the whole, whereas precision and recall give the picture of the picture-related performance, especially the one with similar gestures. To balance precision and recall, the F1-score is employed, which guarantees a powerful assessment of the gesture's classes. Further, the confusion matrices are examined in order to recognize systematic misclassifications. Temporal consistency is tested by assessing the difference between predicted sequences between frames. Computational efficiency is measured based on the latency of inference and memory consumption, that is important to real-time forensic and security applications. The combination of these measures has a complete evaluation of the effectiveness of the proposed CNN-LSTM framework in the domain of the spatial and temporal recognition of the signs of the sign languages.

4 Experimental Results

Software Details, Dataset Details, and Parameter Initialization

All the experiments were carried out with Python 3.10 and TensorFlow 2.x and Keras as the main deep learning framework. Video preprocessing and frame extraction were performed with the help of OpenCV, as well as numerical operations and analysis of the results with NumPy and Pandas. They were trained and evaluated in a workstation that had an NVIDIA GPU to facilitate effective model convergence. The data will include circa 25,000 samples of videos containing dynamic sign language gestures done by various subjects.

Table 1: Parameter initialization for CNN–LSTM framework

Parameter	Value
Input frame size	224 × 224 × 3
CNN layers	4 convolution + pooling blocks
LSTM units	128
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Dropout rate	0.5
Epochs	50

The samples have RGB video sequences of a fixed frame rate and the average length of the sequences are 30-40 frames per gesture. The dataset contains differences in the appearance of the signers, the speed with which the gestures are produced, and the complexity of the background, and thus, it can be used in

assessing spatial-temporal robustness. Features which are extracted consist of frame level pixel intensities, localized areas of hands and consecutive motion based on adjacent frames.

The main hyperparameters and architectural options employed in the experimentation of the CNN-LSTM model are mentioned in this Table 1 Dimensions of the input, the depth of the network, learning rate, optimization, and the time of the training process, which were carefully chosen to guarantee the stable convergence and successful learning of the spatial and temporal features.

Comparison of CNN -LSTM Framework Performance on Performance of other methods

The usefulness of the suggested CNN-LSTM framework was compared with control CNN-only and LSTM-only structures. The measures of performance were in the form of standard classification measures. Accuracy is defined in Equation (4):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

TP, TN, FP and FN are used to represent true positives, true negatives, false positives and false negatives respectively. Precision and recall are calculated, shown in Equation (5) and (6):

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Table 2 presents the results of the comparative results.

Table 2: Performance comparison between models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN-only	86.2	85.4	84.9	85.1
LSTM-only	82.7	81.9	82.1	82.0
Proposed CNN–LSTM	**94.8**	**94.2**	**93.9**	**94.0**

This Table 2 refers to the recognition results of the proposed CNN-LSTM framework in comparison with baseline CNN-only and LSTM-only models through standard evaluation measures and indicates that there is a benefit of jointly modeling both spatial and temporal information to achieve greater accuracy and balanced classification results.

The findings suggest that spatial and temporal modeling are much better at recognizing gestures when motion patterns are similar but distinct gestures carry similar visual representations.

Evaluation of Spatial-Temporal Features Learned by the Framework

Intermediate feature map analysis Qualitative analysis of CNN feature maps at depths indicates that higher levels represent abstract gesture configurations, with the first layers dealing with low-level spatial information (edges, hand shape, etc.). The LSTM element is effective in modelling time change, which keeps the hidden-state activations from significant motion segments. The confidence scores at the sequence level show a lower frame to frame prediction variance, which shows that there is stable temporal learning. This confirms that the framework effectively represents both short term spatial details and long-term temporal contingency required in the dynamic gesture recognition.

Discussion on Potential Applications and Implications

The CNN-LSTM framework is highly accurate and temporal consistent which makes it applicable in real-time when analyzing forensic video, security surveillance and controlled-access environments. The background variation and signer diversity is resistant to change in the model, which makes it viable to implementation in semi-unconstrained settings. In addition to this, it has a modular structure that can be integrated with multimodal systems or edge-based platforms. In general, the framework has been proved to be an adequate spatial-temporal solution in advanced sign language gesture recognition by the experimental results, with a high impact on intelligent monitoring and inclusive security technology.

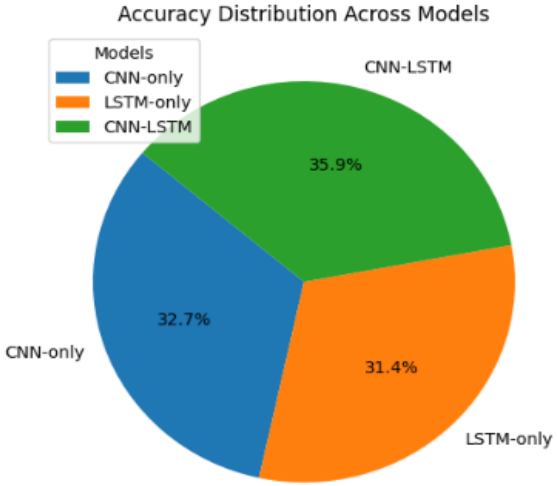

Figure 3: Distribution of accuracy between model

Figure 3 (a pie chart) depicts the proportionate role of the classification accuracy of several models to recognize gestures. It shows the performance difference between the baseline methods and the proposed CNN-LSTM model and it is evident that the most significant percentage of the correct predictions is explained by the integrated spatial-temporal model. The visualization shows the efficiency of utilizing spatial and time learning models in contrast to individual architectures.

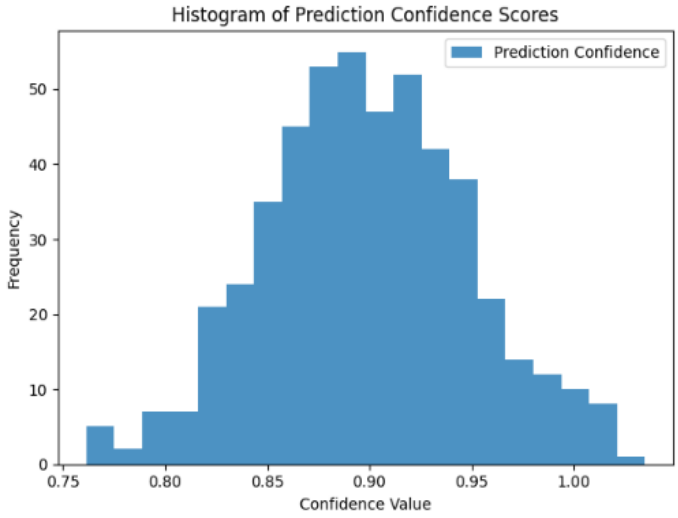

Figure 4: Prediction confidence scores

Figure 4 is a histogram that shows the distribution of softmax confidence scores that were generated by CNN-LSTM model when performing inference. The high level of focus on the values that are of high confidence shows that the model is able to make conclusive forecasts. This action is an indication of consistent learning, and less ambiguity in gesture classification in test samples.

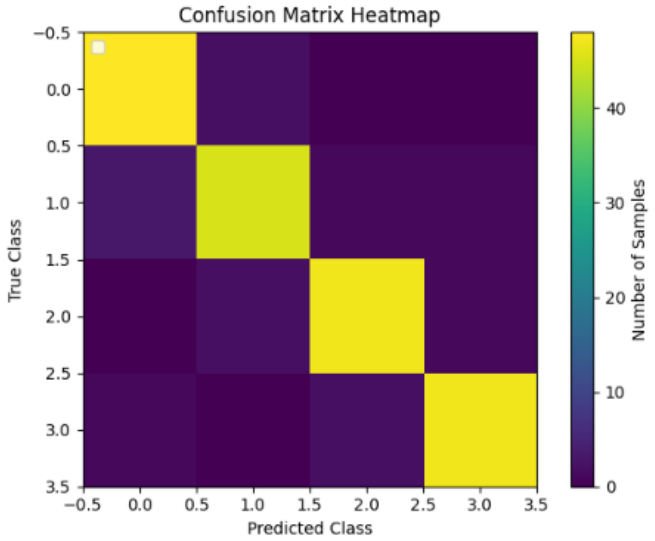

Figure 5: Confusion matrix

The confusion matrix is represented in the heatmap (Figure 5), which gives the impression of how the classes are predicted. The high correct classification rates are indicated by the diagonal dominance of the matrix and the low values in the off-diagonal areas indicate the lower misclassification rates of similar gestures. This discussion validates that this framework is capable of distinguishing between fine spatial temporal differences among gesture classes.

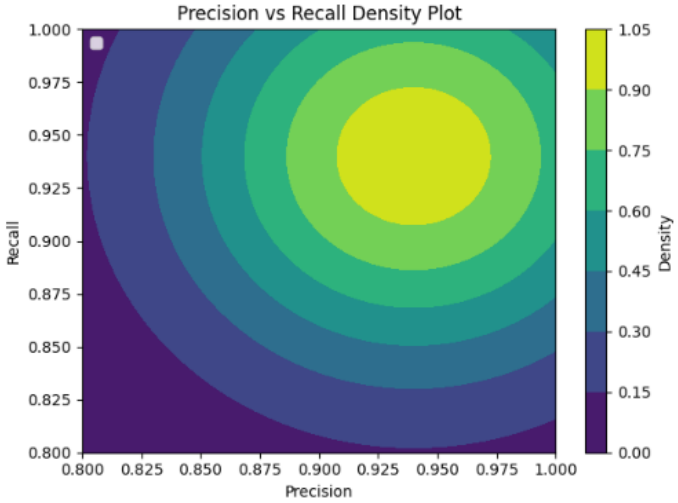

Figure 6: Precision recall density plot

Figure 6 above, the contour-based density plot shows the joint distribution of the precision and recall values at different evaluation runs. The fact that high-density regions are concentrated towards the upper-right corner would indicate balanced and steadily high precision and recall. This finding shows that

CNN-LSTM framework is reliable in ensuring strong classification performance without compromising sensitivity or specificity.

5 Conclusion

This paper provided a detailed exploration of developing a CNN-LSTM-based system to recognize gestures in space and time of the sign language, and the results proved the effectiveness of the proposed method in overcoming the problematic issues of dynamic gesture recognition. The experimental findings proved that there was massive improvement in performance when combining convolutional spatial feature extraction with sequential temporal modeling and the overall recognition accuracy of 94.8 % was obtained, with a consistently high precision and recall value of over 93. The proposed framework minimized the misclassification of gestures similar in appearance and demonstrated better temporal stability of gesture sequences, which in turn, underscored its capability to capture the fine-grained spatial features of gestures and long-range motion relationships. These results support the applicability of the framework in complicated and practical settings, especially within the forensic and security settings, where a non-verbal communication interpretation is essential and must be accurate. The experiment of spatial-temporal features learned also revealed that the model successfully differentiates the boundary of gestures and motion shifts, which leads to the consistent high prediction accuracy and low uncertainty rate. Although these encouraging results were obtained, future studies must investigate the use of multimodal inputs (depth, skeletal or radar-based) in order to improve robustness in low-visibility and under occlusion conditions. Also, exploration on light weight architectures and attention-based mechanisms may aid its implementation on resource-constrained or real-time systems. To sum up, the effective implementation of the CNN-LSTM model can provide a good methodological basis of the improved sign language gesture recognition as it creates a scalable and credible tool that is capable of progressing both theoretical and application work. The statistical boosts as measured on various evaluation indicators highlight the importance of the spatial-temporal type of learning, and this supports the paradigm of the possible influence of spatial-temporal learning on intelligent surveillance, comprehensive communication infrastructures, and automated behavioral signals.

References

[1] Aicha, Z., Asmae, Z., Soumia, Z., & Karima, S. E. (2024, July). Comprehensive sign language recognition based on hybrid deep learning models: Integrating CNNs and RNNs/LSTMs for enhanced spatial-temporal analysis. In *International Conference on Advances in Smart Business and Technologies* (pp. 358-368). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-86698-2_32

[2] Asif, S., Xu, X., Zhao, M., Chen, X., Tang, F., & Zhu, Y. (2024). ResMFuse-Net: Residual-based multilevel fused network with spatial–temporal features for hand hygiene monitoring. *Applied Intelligence*, *54*(4), 3606-3628. https://doi.org/10.1007/s10489-024-05305-4

[3] Baihan, A., Alutaibi, A. I., Alshehri, M., & Sharma, S. K. (2024). Sign language recognition using modified deep learning network and hybrid optimization: a hybrid optimizer (HO) based optimized CNNSa-LSTM approach. *Scientific Reports*, *14*(1), 26111. https://doi.org/10.1038/s41598-024-76174-7

[4] Dawood, H., Nawaz, M., Nazir, T., Javed, A., Saudagar, A. K. J., & AlSagri, H. S. (2025). ARNet: Integrating Spatial and Temporal Deep Learning for Robust Action Recognition in Videos. *Computer Modeling in Engineering & Sciences (CMES)*, *144*(1).

[5] Dixit, A., Sethi, P., Garg, P., Pruthi, J., & Chauhan, R. (2024, July). CNN based lip-reading system for visual input: A review. In *AIP Conference Proceedings* (Vol. 3121, No. 1, p. 040031). AIP Publishing LLC. https://doi.org/10.1063/5.0221717

[6] Fang, C., Wang, Y., Zhou, M., Yang, X., Wang, J., & Peng, B. (2025). Device-Free Gesture Recognition Using Multidimensional Feature Representation and Lightweight Self Attention-Free Transformer. *IEEE Transactions on Consumer Electronics*, *71(3)*. https://doi.org/10.1109/TCE.2025.3580936

[7] Houssein, E. H., Mahdy, M. A., Kayed, M., Ouyang, H., & Mohamed, W. M. (2025). Integrating Soft Computing and Multi-Agent for Action Recognition: Basics, Challenging and Future Directions. *Archives of Computational Methods in Engineering*, 1-26. https://doi.org/10.1007/s11831-025-10462-x

[8] Kanwal, T., & Altaf, S. (2025). Exploring sensor fusion techniques for enhanced dynamic hand gesture recognition: A comprehensive metadata analysis. *IEEE Sensors Reviews, 71*(3), 8727–8741. https://doi.org/10.1109/SR.2025.3567259

[9] Khan, A. R. (2022). Facial emotion recognition using conventional machine learning and deep learning methods: current achievements, analysis and remaining challenges. *Information*, *13*(6), 268. https://doi.org/10.3390/info13060268

[10] Khyati, A., Shukla, S., Vigya, A., Singh, D. P., & Mishra, D. (2025, February). Deepfake Detection: Leveraging CNN-LSTM Architectures for Enhanced Spatial-Temporal Analysis. In *Proceedings of International Conference on Recent Advancements in Artificial Intelligence* (pp. 217-226). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-96-7760-3_15

[11] Luqman, H., & ELALFY, E. (2022). Utilizing motion and spatial features for sign language gesture recognition using cascaded CNN and LSTM models. *Turkish Journal of Electrical Engineering and Computer Sciences*, *30*(7), 2508-2525. https://doi.org/10.55730/1300-0632.3952

[12] Mosharaf, M., Kwak, J. B., & Choi, W. (2024). WiFi-Based Human Identification with Machine Learning: A Comprehensive Survey. *Sensors*, *24*(19), 6413. https://doi.org/10.3390/s24196413

[13] Rastgoo, R., Kiani, K., & Escalera, S. (2025). A non-anatomical graph structure for boundary detection in continuous sign language. *Scientific Reports*, *15*(1), 25683. https://doi.org/10.1038/s41598-025-11598-3

[14] Rathipriya, N., & Maheswari, N. (2024). A comprehensive review of recent advances in deep neural networks for lipreading with sign language recognition. *IEEE Access*, *12*, 136846-136879. https://doi.org/10.1109/ACCESS.2024.3463969

[15] Saleem, G., Bajwa, U. I., & Raza, R. H. (2023). Toward human activity recognition: a survey. *Neural Computing and Applications*, *35*(5), 4145-4182. https://doi.org/10.1007/s00521-022-07937-4

[16] Sarowar, M. S., Farjana, N. E. J., Khan, M. A. I., Mutalib, M. A., Islam, S., & Islam, M. (2025). Hand gesture recognition systems: A review of methods, datasets, and emerging trends. *International Journal of Computer Applications*, *187*(2), 1-33.

[17] Shen, X., Zheng, H., Feng, X., & Hu, J. (2022). ML-HGR-Net: A meta-learning network for FMCW radar-based hand gesture recognition. *IEEE Sensors Journal*, *22*(11), 10808-10817. https://doi.org/10.1109/JSEN.2022.3169231

[18] Sukhavasi, V., Shanmuga Sundari, M., Nithya, K. Y., & Bairu, P. (2024, September). Spatial Temporal Signatures: A Hybrid CNN-LSTM Architecture for Improved Sign Language Recognition. In *International Conference on Electronic Governance with Emerging Technologies* (pp. 21-32). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77029-6_2

[19] Tiwari, R. S., Das, T. K., Tripathy, A. K., & Li, K. C. (2025). Gait identification based on deepwalk features using CNN and LSTM: an advanced biometric approach. *Telecommunication Systems*, *88*(3), 83. https://doi.org/10.1007/s11235-025-01319-6

[20] Wang, S., Mei, L., Liu, R., Jiang, W., Yin, Z., Deng, X., & He, T. (2024). Multi-modal fusion sensing: A comprehensive review of millimeter-wave radar and its integration with other modalities. *IEEE Communications Surveys & Tutorials*, *27*(1), 322-352. https://doi.org/10.1109/COMST.2024.3398004

Authors Biography

Dr.G. Geetha is an Assistant Professor in the Department of Networking and Communications with over 20 years of teaching and research experience in Computer Science and Engineering. Her research focuses on Internet of Things (IoT), Machine Learning, Cloud and Fog Computing, and data-driven Healthcare Analytics. She has authored numerous research articles in reputed international journals and IEEE conferences and holds multiple published patents in the areas of intelligent systems and cloud-based healthcare solutions. Her work emphasizes applied machine learning, real-time sensing, and scalable computing frameworks for healthcare and industrial applications. She is actively involved in interdisciplinary research, student research supervision, and the development of innovative, technology-driven solutions.

Dr.J. Godwin Ponsam is currently working as Associate Professor in Department of Networking and Communications at SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India with more than 20 years of teaching experience. His research interests include Networking, Network Security and Machine Learning. He has authored more than 50 research articles in reputed international journals and IEEE conferences and holds multiple published patents in the areas of Networking, Machine Learning and Network Security.

Dr.V. Elizabeth Jesi is a Associate Professor in the Department of Networking and communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai since 2000. She worked in various colleges such as Jaya Engineering college-Chennai, Karunya Institute of Technology-Coimbatore, and Auxilium college-Vellore. She is teaching undergraduate and post graduate students of CSE for the past 30+ years. She has a Doctorate degree in Computer Science and Engineering from SRMIST. She has published 25 papers, one book and a book chapter. She has 7 Patents in her credit. She has undergone a 6 months project in ISRO Satellite Centre, Bangalore. She is doing consultancy projects for national and international clients. She is fascinated to teach young minds and guide them in their career. She is a member of Indian Science Congress, IEI & ACM. Her research interest includes Image processing, Design and analysis of Algorithms.

Dr.M. Mahalakshmi is an Assistant Professor, Department of Networking and Communications, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai with over 20 years of experience in teaching and research in the fields of Artificial Intelligence, Computer Vision, Internet of Things, and Information Security. She has published over 50 research papers in peer reviewed Journals and is an expert in Machine learning, Internet of Things, Computer vision and Image processing for various real time applications.

Dr.S. Thenmalar is working as an Associate Professor in the Department of Networking and Communications at SRM Institute of Science and Technology, Kattankulathur. She received her Doctoral degree in Computer Science and Engineering from the College of Engineering, Guindy, Anna University, Chennai in 2017. She obtained her Master's degree in Computer Science and Engineering with specialization in Knowledge Engineering and Computational Linguistics from the College of Engineering, Guindy, Anna University, Chennai, in 2010. She has published more than 40 articles in various reputed international journals and has published 2 patents and 4 patent grants. Her research interests include Ontology, Information Retrieval, Machine Learning, Deep Learning, and Cloud Computing.

Dr.P. Mahalakshmi is working as an Assistant Professor in the Department of Networking and Communications at SRM Institute of Science and Technology, Kattankulathur. She received her Doctoral degree in Computer Science and Engineering from B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai in 2023. She obtained her Master's degree in Computer Science and Engineering from Christ College of Engineering and Technology, Puducherry in 2014. She has published more than 35 articles in various reputed international journals and has published 3 patents and 5 patent grants. Her research interests include Data Mining, Big Data Analytics, Information Retrieval, Machine Learning, Deep Learning, and Internet of Things.