HySTA-NET : A HYBRID SPATIO-TEMPORAL ATTENTION NETWORK FOR RECOGNITION OF EMERGENCY SIGNS IN INDIAN SIGN LANGUAGE

A Project Report
Submitted to the APJ Abdul Kalam Technological University
in partial fulfillment of requirements for the award of degree

Bachelor of Technology
in
Computer Science and Engineering
by

AISHA THAMEEM (TKM21CS012)

BHAGYA A JAI (TKM21CS041)

FATHIMA A (TKM21CS053)

UTHARA SABU (TKM21CS138)

Department of Computer Science and Engineering
T.K.M College of Engineering, Kollam
April 2025

DECLARATION

We hereby declare that the project report **'HySTA-Net: A HYBRID SPATIO-TEMPORAL ATTENTION NETWORK FOR RECOGNITION OF EMERGENCY SIGNS IN INDIAN SIGN LANGUAGE RECOGNITION'**, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by us under supervision of Dr. Shyna A. This submission represents our ideas in our own words and where ideas or words of others have been included, we have adequately and accurately cited and referenced the

We also declare that we have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not previously formed the basis for the award of any degree, diploma, or similar title of any other University.

Aisha Thameem
Bhagya A Jai
Fathima A
Uthara Sabu

Place: Kollam
Date:28/03/2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING TKM COLLEGE OF ENGINEERING, KOLLAM

CERTIFICATE

This is to certify that the report entitled **'HySTA-Net: A HYBRID SPATIO-TEMPORAL ATTENTION NETWORK FOR RECOGNITION OF EMERGENCY SIGNS IN INDIAN SIGN LANGUAGE RECOGNITION'** submitted by **Aisha Thameem (TKM21CS012), Bhagya A Jai (TKM21CS041), Fathima A (TKM21CS053), Uthara Sabu (TKM21CS138)** to the APJ Abdul Kalam Technological University in partial fulfillment of the B.Tech. degree in Computer Science and Engineering is a bonafide record of the **project** work carried out by them under our guidance and supervision. This report in any form has not been submitted to any other University or purpose.

Project Guide **Head of the Department**

Internal Supervisor **External Examiner**

ACKNOWLEDGEMENT

We take this opportunity to express our deep sense of gratitude to the Almighty and extend our sincere thanks to everyone who helped us complete this project successfully.

We would like to express our sincere gratitude to **Dr. Sajeeb R**., Principal, TKMCE, for providing us with all the necessary facilities and support for carrying out this project. We are extremely grateful to **Dr. Aneesh G. Nath**, Associate Professor and Head of Department, Department of Computer Science and Engineering, **Dr. Ansamma John**, Project Coordinator and Professor, Department of Computer Science and Engineering, and **Dr. Manu J. Pillai**, Associate Professor, Department of Computer Science and Engineering, for their constructive guidance, advice, constant support, and technical guidance provided throughout the development of this project. Without their intellectual support and timely suggestions, this project would not have been possible.

We would also like to express our sincere gratitude to our project guide, **Dr. Shyna A**, Assistant Professor, Department of Computer Science and Engineering, TKM College of Engineering, Kollam, for providing us with all the necessary facilities, guidance, and support throughout the project.

Our immense gratitude extends to all faculty members and technical staff in the Department of Computer Science and Engineering for their assistance and for providing us with the necessary facilities to complete the project. Finally, we thank our families and friends, whose encouragement and support contributed significantly to the successful fulfillment of this project.

Aisha Thameem
Bhagya A Jai
Fathima A
Uthara Sabu

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Communication is essential for all human beings to fulfill their needs and engage with others. For the hearing and speech-impaired community, sign language plays a crucial role in expressing thoughts, accessing education, securing employment, and asserting societal rights. However, a communication gap persists between the general population and the speech-impaired community due to limited training and awareness in sign language, often resulting in misunderstandings, social isolation, and restricted access to essential services. This communication barrier becomes even more critical in emergency situations, where the inability to convey urgent messages can lead to life-threatening consequences.

To address this challenge, this project proposes **HySTA-Net (Hybrid Spatio-Temporal Attention Network)** for Indian Sign Language (ISL) recognition, focused on eight emergency signs. These signs are dynamic in nature, involving hand movements over time, and are captured in video format. Frames were extracted from the videos and processed for classification using four deep learning-based architectures: HySTA-Net using VGG16, HySTA-Net using ResNet50, HySTA-Net using InceptionV3+LSTM, and HySTA-Net using 3D CNN. The classification performance of each model was evaluated, with accuracy, precision, recall, and F1-score serving as key performance metrics. Among these, HySTA-Net with VGG16+LSTM demonstrated the highest classification accuracy.

Additionally, a custom dataset was created to test the models' generalisation ability on new, unseen videos. This dataset consisted of approximately 50 sign videos per gesture, recorded with different signers to account for variations in hand shape, motion speed, and background conditions. The models' performance on this dataset was also assessed to ensure robustness. The results indicate that the proposed system can serve as a reliable and efficient tool for real-time emergency sign recognition, significantly improving accessibility and responsiveness for the speech and hearing-impaired community in critical situations.

Chapter 1

Introduction

Globally, over 466 million people live with hearing loss, and a significant portion of these individuals rely on sign language as their primary means of communication. Sign language serves as a vital mode of communication for individuals who are deaf or hard of hearing, enabling them to convey thoughts, emotions, and urgent needs through structured hand movements, facial expressions, and body gestures. Among various sign languages, Indian Sign Language (ISL) is predominantly used in India. However, its recognition systems, particularly for emergency-related gestures, remain underdeveloped. In critical situations such as medical emergencies, accidents, or disasters, the inability to communicate distress signals effectively can lead to delayed assistance and potentially life-threatening consequences.

Despite advancements in sign language recognition, most existing systems emphasize general vocabulary rather than emergency-specific gestures. Recognizing emergency ISL gestures in real-time presents unique challenges, including gesture variability, self-occlusion, spatio-temporal complexity, and environmental factors such as lighting and background noise. Addressing these complexities requires a robust system capable of accurately interpreting dynamic gestures across different signers and real-world conditions.

The goal of the work is to develop an efficient real-time ISL emergency sign recognition system capable of accurately classifying and translating emergency gestures into readable text. By leveraging deep learning architectures such as CNNs, LSTMs, and multi-head attention mechanisms, the proposed system enhances accessibility and responsiveness, ensuring that no distress signal goes unnoticed.

Developing such a real-time recognition system introduces several technical challenges due to the dynamic nature of gestures and user variability. Some of the key challenges include:

- Self-Occlusion: Complex hand movements often result in partial or complete occlusion, where parts of the hands or fingers block each other, making it difficult to capture complete gesture information.
- Gesture Variability: Different signers perform the same gesture with variations in speed, angles, and intensity, causing inconsistencies in recognition. Differences in hand size, movement trajectory, and orientation further complicate the task.
- Spatio-Temporal Complexity: Unlike static gestures, emergency ISL signs involve continuous hand movements over time, necessitating models that effectively capture both spatial and temporal features.
- Background Noise and Lighting Variations: Real-time applications must function reliably across diverse environments with varying lighting conditions, backgrounds, and potential motion blur.
- Limited Data Availability: Emergency ISL datasets are scarce, making it necessary to employ data augmentation or synthetic data generation techniques to improve generalization.
- Similar Gestures with Different Meanings: Subtle differences in hand positioning and motion may carry distinct meanings, demanding precise feature extraction.

Overcoming these challenges requires a robust system that handles such variations in real-time. The proposed approach utilizes deep learning models integrating convolutional and recurrent architectures for efficient spatio-temporal feature extraction. A custom dataset is also employed to improve generalization across different signers and environments, ensuring that emergency gestures are accurately recognized in real-world conditions. The framework aims to enhance communication for the deaf and hard-of-hearing community, particularly in critical situations where timely assistance is essential.

1.1 Motivation

Effective emergency communication is crucial for ensuring timely assistance, yet deaf and hard-of-hearing individuals often face significant challenges in conveying urgent messages due to the lack of widespread sign language literacy. In high-stakes situations such as medical

emergencies, accidents, or disasters, delays in understanding ISL gestures can have life-threatening consequences.

While advancements in sign language recognition have improved accessibility in general contexts, specialised systems for emergency gestures remain underdeveloped. Real-time recognition of emergency gestures involves dynamic movements that vary across individuals and environments, posing considerable technical challenges.

The work addresses the need for a reliable emergency ISL recognition system by leveraging deep learning models that can interpret critical gestures in real-time, ensuring timely response and improved safety. By empowering the deaf and hard-of-hearing community, the proposed system enhances accessibility and independence, ensuring no call for help goes unnoticed in critical situations.

1.2 Problem Statement

In life-threatening situations, timely communication is paramount. However, deaf and hard-of-hearing individuals face significant barriers when conveying emergency messages, as most emergency response systems rely on verbal or text-based communication. The absence of an Indian Sign Language (ISL) recognition system for emergency scenarios creates a dangerous gap, making it challenging for sign language users to seek immediate help during:

- Medical Emergencies (e.g., heart attacks, injuries)
- Accidents and Disasters (e.g., road accidents, fires)
- Threats and Crimes (e.g., theft, assault)
- Public Spaces and Transportation (e.g., airports, trains)

Despite advancements in artificial intelligence and gesture recognition, existing systems primarily focus on general vocabulary and lack support for emergency-specific signs. This gap prevents deaf individuals from accessing rapid assistance in urgent situations, compromising their safety and well-being.

The goal of the work is to develop a real-time ISL emergency sign recognition system that accurately interprets distress gestures and converts them into readable text or speech. By leveraging deep learning models, including CNNs, LSTMs, and multi-head attention mechanisms, the system ensures efficient and precise recognition of emergency gestures, facilitating seamless communication between sign language users and emergency responders.

1.3 Objectives

The proposed work aims to develop an efficient real-time Indian Sign Language (ISL) emergency gesture recognition system to enhance communication for deaf and hard-of-hearing individuals in critical situations. The key objectives are outlined as follows:

- Enabling Real-Time Recognition of Emergency ISL Gestures: To implement a system capable of accurately recognizing eight dynamic emergency gestures in real-time, addressing variations in signing speed, hand movements, and environmental conditions.

- Developing Deep Learning-Based Classification Models: To design and optimize classification models, including CNN-LSTM architectures, for effective recognition of dynamic gestures by capturing both spatial and temporal features.

- Evaluating System Performance with Quantitative Metrics: To assess the model's effectiveness using metrics such as accuracy, precision, recall, and F1-score, along with a confusion matrix to reduce misclassifications.

- Developing a Custom Video Dataset for Comprehensive Evaluation: To create a specialized test dataset of emergency ISL gestures featuring diverse signers, backgrounds, and lighting conditions, allowing thorough performance assessment and improving the model's generalization ability.

- Enhancing Accessibility in Emergency Situations: To provide a practical and scalable solution that enables deaf and hard-of-hearing individuals to communicate distress signals effectively, ensuring timely response and improving safety in critical scenarios.

1.4 Organization of Report

The report begins with an extensive literature review in Chapter 2, providing foundational insights and related research studies. Chapter 3 outlines the methodologies applied, detailing the specific approaches taken to address project objectives. and also elaborates the project implementation process, covering each development phase in detail. Chapter 4 presents the testing and validation procedures, demonstrating how model performance was evaluated. Finally, Chapter 5 offers a comprehensive conclusion, summarizing key findings, discussing the project's impact, and proposing future directions. The report concludes with a list of bibliographical references, acknowledging the sources that contributed to this work.

Chapter 2

Literature Survey

Early studies in hand gesture recognition primarily focused on static gestures using Convolutional Neural Networks (CNNs) to perform classification and detection tasks. Researchers demonstrated that CNNs could achieve cutting-edge performance on image-based recognition tasks. For example, early implementations successfully applied CNNs to classify static ASL gestures with high accuracy; however, these approaches were inherently limited when extended to dynamic gestures since they lacked the capability to model temporal dependencies.

Subsequent work addressed this limitation by integrating temporal modeling into the recognition pipeline. Rao et al. introduced a neural network classifier for continuous sign language identification using selfies, while later studies employed recurrent structures to capture the sequential nature of dynamic gestures. Notably, Do et al.[2] proposed a multi-level feature LSTM network on a dynamic hand gesture dataset, achieving accuracies of 96.07% and 94.40% for 14 and 28-class problems, respectively. Despite the high performance, these methods often struggled with limited datasets and the complexity of modeling long-term dependencies.

Building on these advances, Cui et al.[3] developed a video-based recurrent CNN to address continuous sign language recognition, and Elboushaki et al.[4] combined Residual Networks (ResNets) with Convolutional LSTM (ConvLSTM) to capture spatio-temporal interdependencies. Liao et al.[5] further refined this approach by proposing a multimodal dynamic recognition system based on a 3D residual ConvNet coupled with bi-directional LSTM networks (B3D ResNet), which effectively handled complex hand gestures but still faced challenges related to dataset limitations and computational complexity.

In parallel, researchers also explored the integration of frame extraction techniques to enhance dynamic recognition. John et al.[6] demonstrated that efficient extraction of representative frames from video sequences could significantly improve the performance of long-term recurrent convolutional networks (LRCN). Similarly, Lai and Yanushkevich[7] combined depth and skeleton data using CNNs and RNNs, achieving an overall accuracy of 85.46%, though the system's reliance on additional sensor data posed practical constraints.

Further advancements incorporated multi-modal approaches and advanced CNN architectures. Obaid et al.[8] proposed a two-stage model for hand gesture recognition in video sequences, leveraging CNN and RNN-LSTM architectures to label and classify frames. While this method improved performance, it still encountered challenges in capturing the full complexity of dynamic gestures. Additional works by Molchanov et al.[9] combined 3D CNNs and RNNs to enhance gesture recognition, and Saqib et al.[10] and Camgoz et al.[11] successfully employed 3D CNNs to capture discriminative spatio-temporal features, achieving accuracies above 90%. However, the high computational cost of 3D CNNs remained a drawback.

Object detection frameworks, such as YOLO v5 combined with DarkNet-53[12], have also been applied to hand gesture detection in complex environments. These models achieved remarkable detection accuracy (up to 97.68%), yet their application is often restricted to low-resolution images and simpler scenarios. In the realm of static gesture recognition, Khari et al. [13] proposed a method using a pre-trained VGG19 model on the ASL dataset, achieving an accuracy of 94.8%. While such models and specialized CNN architectures have demonstrated impressive performance, they are less effective in modeling the temporal evolution necessary for dynamic gestures.

More recent approaches have integrated pre-trained CNNs, such as VGG-16 or Inception-v3, with LSTM networks to effectively model dynamic gestures. Masood et al.[14] combined Inception-v3 with various LSTM units for the Argentinean Sign Language dataset,

attaining a best accuracy of 95.2%. Adithya and Rajesh[15] further explored this strategy by employing both a pre-trained GoogleNet with LSTM and a multi-class SVM, highlighting the potential of transfer learning for dynamic sign language recognition.

ISL presents additional challenges compared to ASL, as it involves bilateral hand gestures and more complex, dynamic movements. The limited availability of benchmark datasets for ISL has further hindered progress in this area. To address these challenges, the work proposes a comprehensive model that integrates advanced deep learning techniques, combining pre-trained VGG-16 for spatial feature extraction, LSTM for temporal modeling, and multi-head attention with residual connections to effectively recognize emergency ISL signs from video data. This model is designed to overcome the limitations of previous approaches by capturing both spatial and temporal features with high accuracy, thus offering a robust solution for emergency communication among the deaf community.

Chapter 3

Methodology

Sign language recognition plays a vital role in breaking communication barriers for the deaf and hard-of-hearing community, facilitating accessibility in various domains, including healthcare, emergency response, and daily life. Accurate classification of emergency gestures in ISL is crucial for enabling swift and effective communication, particularly in critical situations where immediate response is essential.

A novel architecture, entitled **"Hybrid Spatio-Temporal Attention Network (HySTA-Net)"**, is introduced for the recognition of emergency Indian Sign Language gestures. The comprehensive framework integrates a series of specialised modules designed to efficiently process video data and extract meaningful features from both spatial and temporal domains. A detailed diagram of the architecture is provided in Figure 1, illustrating the sequential data flow through each module and emphasising the technical rationale behind every design choice, ultimately offering an innovative and effective approach to emergency ISL recognition.

The system begins with a Video Acquisition Module that captures raw video streams, followed by a Pre-Processing Module where frames are standardised through resizing, normalisation, and contrast enhancement. Subsequently, the Spatial Feature Extraction Module leverages a modified VGG-16 network to derive high-level spatial features, which are then passed to a Temporal Feature Extraction Module employing LSTM to capture dynamic motion patterns across time.

Building on these foundational features, the architecture incorporates a Multi-Layered Attention Module that refines the temporal representations by attending to critical aspects of the gesture sequence. The output from the attention mechanism is further aggregated via a Global Average Pooling Module to produce a concise feature vector, which is then processed by a Classification Module comprising dense layers with non-linear activations and dropout regularisation. Finally, the Output Module interprets the softmax probabilities to yield the final gesture recognition results.

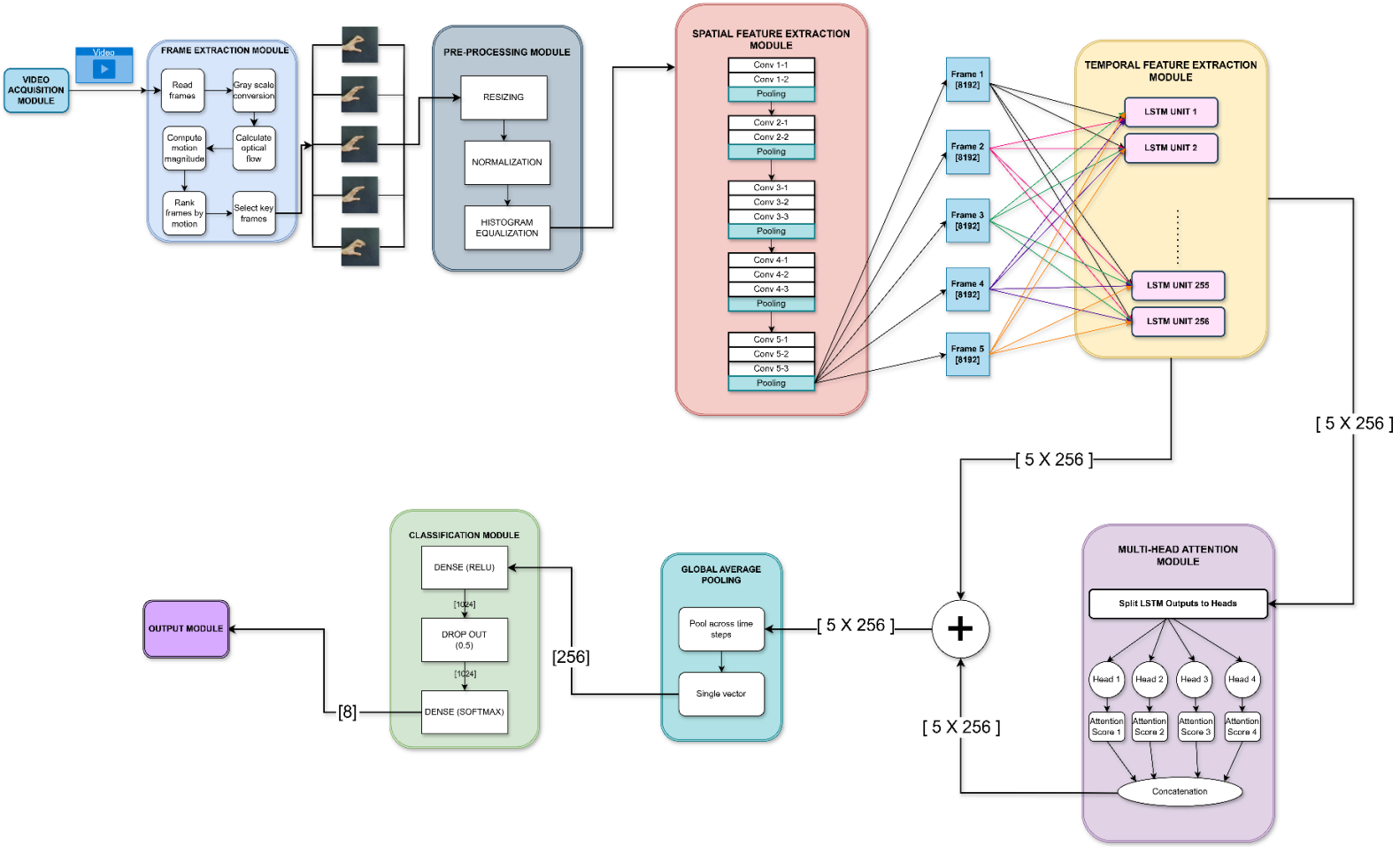


Figure 1: Proposed Architecture Diagram of HySTA-Net

3.1 Video Acquisition Module

The Video Acquisition Module is responsible for providing the raw video data to the subsequent Frame Extraction Module. In the work, a total of 412 video samples have been collected, representing 8 distinct emergency sign classes in Indian Sign Language. These videos, stored in a structured directory format with each class organized in separate sub-folders, serve as the primary data source for the system. The input module ensures that the video data is reliably accessed, properly organized, and ready for further preprocessing. This initial step lays the foundation for effective frame extraction and subsequent feature analysis.

This module seamlessly transitions to the next stage, where the Frame Extraction module processes the input videos to capture the most significant motion cues, as will be discussed in the following section.

3.2 Frame Extraction Module

Given that the input videos are dynamic and not all frames contain useful information, selecting keyframes is crucial for reducing redundancy while preserving essential motion cues for accurate recognition. To achieve this, the Frame Extraction Module systematically converts dynamic video sequences into a structured set of key frames for analysis.

Various frame extraction methods were evaluated such as uniform sampling, random selection, and other motion-based approaches. Through experimentation, the hybrid optical flow method emerged as the most effective, as it quantifies the motion between consecutive frames and selects only those exhibiting significant movement, ensuring that the beginning and ending of the gesture are always captured. This approach effectively distinguishes key movement moments, providing the model with a consistent and informative input that accurately reflects gesture dynamics. The resulting output, a set of high-quality, motion-rich frames forms the basis for further processing in the Preprocessing Module.

Algorithm Overview:

Input: A dynamic video sequence containing emergency sign language gestures.

Output**:** A curated set of frames that effectively represents the dynamic gesture, ready for preprocessing (resizing, normalization) in the next module.

1. Motion Quantification:
    - Compute the optical flow between consecutive frames using the Farneback method.
    - Calculate the motion magnitude using:

$$\|\mathbf{f}\| = \sqrt{u^2 + v^2}$$

where *u* and *v* are the horizontal and vertical flow components, respectively.

2. Key Frame Selection:
    - Identify frames with the highest motion magnitudes (motion peaks).
    - Ensure that the first and last frames are always included.
    - If the total number of selected frames is less than the required sequence length, supplement with uniformly sampled frames.

3.3 Pre-Processing Module

The Pre-processing Module standardizes and enhances the extracted frames to ensure optimal performance of the subsequent spatial-temporal analysis using VGG-16. This module performs key operations such as resizing, normalization, and histogram equalization to refine the input data. Each frame is resized to a consistent dimension (e.g., 150×150 pixels) to maintain uniformity across the dataset. Normalization is then applied, scaling pixel intensity values to a

range of [0,1], which harmonizes the data distribution and stabilizes model convergence during training. To further enhance contrast and highlight subtle gesture features, histogram equalization is performed. After these transformations, the pre-processed frames are converted into a consistent format and fed into the spatial-temporal module (VGG-16) for feature extraction and recognition.

3.4 Spatial Feature Extraction Module

The Spatial Feature Extraction Module is designed to derive high-level, discriminative representations from each pre-processed frame. These spatial features are critical for capturing the static and structural details of emergency sign language gestures. Initially, multiple architectures were evaluated for spatial feature extraction, including a 3D Convolutional Neural Network (3D-CNN), Inception V3, and ResNet50. However, limitations such as the relatively small dataset size and increased computational overhead led to suboptimal performance with these models. In contrast, VGG-16 demonstrated superior capability in extracting fine-grained spatial features, making it the preferred choice for our application.

VGG-16's architecture, characterized by its deep stack of small 3×3 convolutional filters, effectively captures hierarchical spatial representations. In this implementation, the top fully connected layers of VGG-16 have been removed, and the convolutional base is frozen to leverage the pre-trained weights on ImageNet. This transfer learning approach minimizes overfitting while enabling robust feature extraction despite limited training data.

For each pre-processed frame, the VGG-16 network extracts a robust set of spatial feature maps that capture essential details such as edges, textures, and complex structures inherent to the gesture images. Extensive experimentation showed that VGG-16 outperforms 3D-CNN, Inception V3, and ResNet-15 in our scenario, thereby providing a more discriminative feature representation.

Once these spatial features are extracted, they are arranged into a sequence of feature vectors that encapsulate the spatial characteristics of the input video. This output is then

seamlessly fed into the Temporal Feature Extraction Module, where the dynamic temporal dependencies of the gestures are modeled further.

3.5 Temporal Feature Extraction Module

Temporal Feature Extraction is essential for capturing the dynamic evolution of sign language gestures—information that static spatial features alone cannot convey. Given the inherently sequential nature of dynamic gestures, modeling temporal dependencies is crucial for accurate recognition.

To address this, we employ a Long Short-Term Memory (LSTM) network with 256 hidden units. The LSTM is well-suited for sequential data as it maintains an internal memory state that can capture both short-term transitions and long-term dependencies. For an input sequence of spatial feature vectors $x_1, x_2, \ldots, x_T$ (where each x_t is derived from the VGG-16 module), the LSTM processes the sequence iteratively as follows:

$$h_t = \text{LSTM}(x_t, h_{t-1})$$

where:

- x_t is the input feature vector at time t,
- h_{t-1} is the previous hidden state,
- h_t is the updated hidden state at time t.

Each of the 256 units in the LSTM adapts its internal state to capture the temporal context of the gesture sequence. The output of the LSTM is a sequence of hidden states $h_1, h_2, \ldots, h_T$ that encapsulate the dynamic progression of the sign language gesture.

This module, by leveraging a 256-unit LSTM, effectively encodes the time-dependent behavior of dynamic gestures, serving as a critical step in our overall emergency sign language recognition system.

After processing through the LSTM block, the resultant temporal features provide a comprehensive representation of the gesture's evolution. These features are then passed to the subsequent module for further refinement. In particular, additional mechanisms such as multi-head attention may be applied later to further enhance the model's focus on the most salient temporal cues.

3.6 Multi-head Attention Module

The use of a multi-head attention mechanism is proposed to further enhance the temporal feature representation obtained from the LSTM. Unlike single-head attention, which computes a single set of attention weights and captures only one aspect of the temporal dependencies, multi-head attention enables the model to attend to information from multiple representation subspaces concurrently. This is particularly advantageous in dynamic sign language recognition, where capturing diverse aspects of motion and context is crucial.

In this approach, the attention mechanism divides the LSTM output into multiple subspaces. For each head, the scaled dot-product attention is computed as:

$$\text{Attention(Q,K,V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where:

- Q (queries), K(keys), and V(values) are linear projections of the LSTM output,
- d_k is the dimensionality of the key vectors,
- The *softmax* function normalizes the attention weights.

Assuming the model utilizes four parallel attention heads, the outputs from each head are concatenated and then linearly transformed:

$$\text{MultiHead}(Q,\ K,\ V)\ =\ \text{Concat}(head_1, head_2, head_3,\ head_4)W^O$$

where W^O is the output projection matrix. This mechanism allows the model to capture a richer set of features by attending to multiple aspects of the temporal data simultaneously.

The enhanced representation provided by the multi-head attention module is expected to improve the model's ability to discern subtle temporal patterns, thereby contributing to more accurate recognition of dynamic emergency sign language gestures. The output of this module is subsequently passed to further processing stages in the overall architecture.

3.7 Residual Connection

The use of a residual connection is justified by the need to preserve the original temporal features extracted by the LSTM while integrating the enhanced contextual information from the multi-head attention module. In dynamic sign language recognition, it is critical that the original sequential signal is maintained; the residual connection facilitates this by providing a direct pathway that bypasses the attention refinement. This not only helps in mitigating issues like vanishing gradients during training but also allows the network to learn an identity mapping, ensuring that the essential temporal dynamics are not lost. In our architecture, the residual connection is implemented as an element-wise addition of the LSTM output (x_{LSTM}) and the output from the multi-head attention module (x_{attn}). This operation is mathematically represented as:

$$x_{res} = x_{LSTM} + x_{attn}$$

Here:

- x_{LSTM} represents the temporal feature vector output from the LSTM module.

- x_{attn} denotes the refined feature vector from the multi-head attention module.
- x_{res} is the resulting feature vector after the residual addition.

This summative operation ensures that both the original and refined features contribute to the final representation, thereby enhancing the overall discriminative capability of the model. The improved feature representation aids in the accurate recognition of dynamic emergency sign language gestures, meeting the objectives.

3.8 Global Average Pooling

The Global Average Pooling (GAP) Module is employed to aggregate the temporal feature representations into a single, fixed-length vector. This module pools across the time dimension of the input feature sequence, which can be mathematically expressed as:

$$z = \frac{1}{T}\sum_{t=1}^{T} x_t$$

where:

- x_t is the feature vector at time step t
- T is the total number of time steps,
- z is the resulting aggregated feature vector.

The significance of GAP lies in its ability to reduce the spatial-temporal feature map into a concise representation, which not only minimizes the number of parameters but also helps mitigate overfitting. By averaging across time steps, GAP preserves the most salient features required for accurate gesture recognition and prepares the data for the final classification.

3.9 Classification Module

The Classification Module converts the pooled feature vector into a probability distribution over the sign classes. This module comprises the following layers:

1. **Dense Layer with ReLU Activation:**
   This layer transforms the aggregated feature vector into a higher-dimensional space and introduces non-linearity via the ReLU (Rectified Linear Unit) activation function. The operation can be represented as:

$$h = \text{ReLU}(W_1 z + b_1)$$

   where:

   - z is the input feature vector from the GAP module,
   - W_1 and b_1 are the weight matrix and bias vector for this layer,
   - h is the output of the dense layer after applying ReLU.

2. **Dropout Layer:**
   A dropout layer with a dropout rate of 0.5 is applied to prevent overfitting by randomly deactivating 50% of the neurons during each training iteration. This regularization technique encourages the network to learn more robust features.

3. **Dense Layer with Softmax Activation:**
   Finally, the output from the dropout layer is fed into a dense layer that uses the softmax activation function to produce the final probability distribution over the classes:

$$y = \text{softmax}(W_2 h + b_2)$$

where:

- W_2 and b_2 are the weight matrix and bias vector for this layer,
- y is the output vector containing the probabilities for each sign class.

The classification module thus effectively transforms the refined, pooled features into a set of predictions that indicate the likelihood of each emergency sign.

3.10 Output Module

The Output Module captures the final predictions of the system. The softmax output vector *y* represents the probability distribution over the predefined sign classes. The class with the highest probability is selected as the final predicted sign. This output is crucial for the evaluation and deployment of the sign language recognition system, providing a clear and interpretable decision that can be used in real-time applications.

3.11 Custom Models created as part of Methodology

As part of this study, multiple deep learning architectures were designed and evaluated for dynamic ISL gesture recognition.The following subsections provide a brief description of the architectures explored during this study. Among these, HySTA-Net using VGG16 as the spatial feature extractor demonstrated the best performance in both train-test splits and evaluations on custom video datasets

- HySTA-Net using Resnet50

  In this architecture, ResNet50 was employed as the spatial feature extractor due to its deep residual learning framework, which allows for effective gradient propagation and improved feature representation. The extracted spatial features were then passed to an

LSTM-based temporal feature extraction module, enabling the model to capture sequential dependencies in gesture movements.

Instead of a multihead attention mechanism, this model incorporated a self-attention mechanism, as it demonstrated improved accuracy and better feature refinement in this configuration. The self-attention module helped the model focus on the most relevant temporal features, enhancing its ability to recognize dynamic gestures effectively.

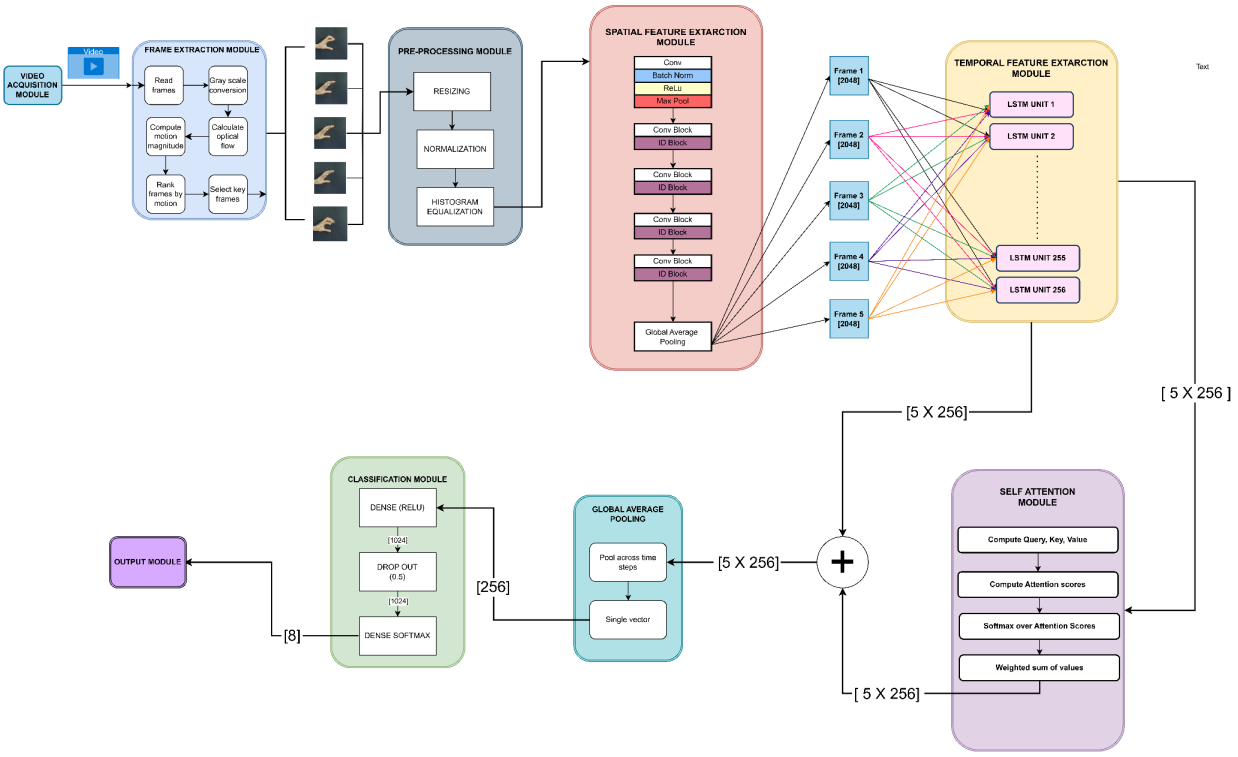


Figure 2: HySTA-Net using Resnet50

- HySTA-Net using InceptionV3

In this model, InceptionV3 was utilized as the spatial feature extraction module due to its efficient architectural design, which incorporates multiple filter sizes within the same

layer to capture spatial features at different scales. The extracted spatial features were then passed through an LSTM-based temporal feature extraction module, allowing the model to learn sequential dependencies in gesture movements.

The multi-head attention mechanism incorporated into this model refines the extracted temporal features by focusing on different parts of the sequence simultaneously. This attention mechanism enhanced the model's ability to capture intricate motion patterns, improving gesture recognition performance.

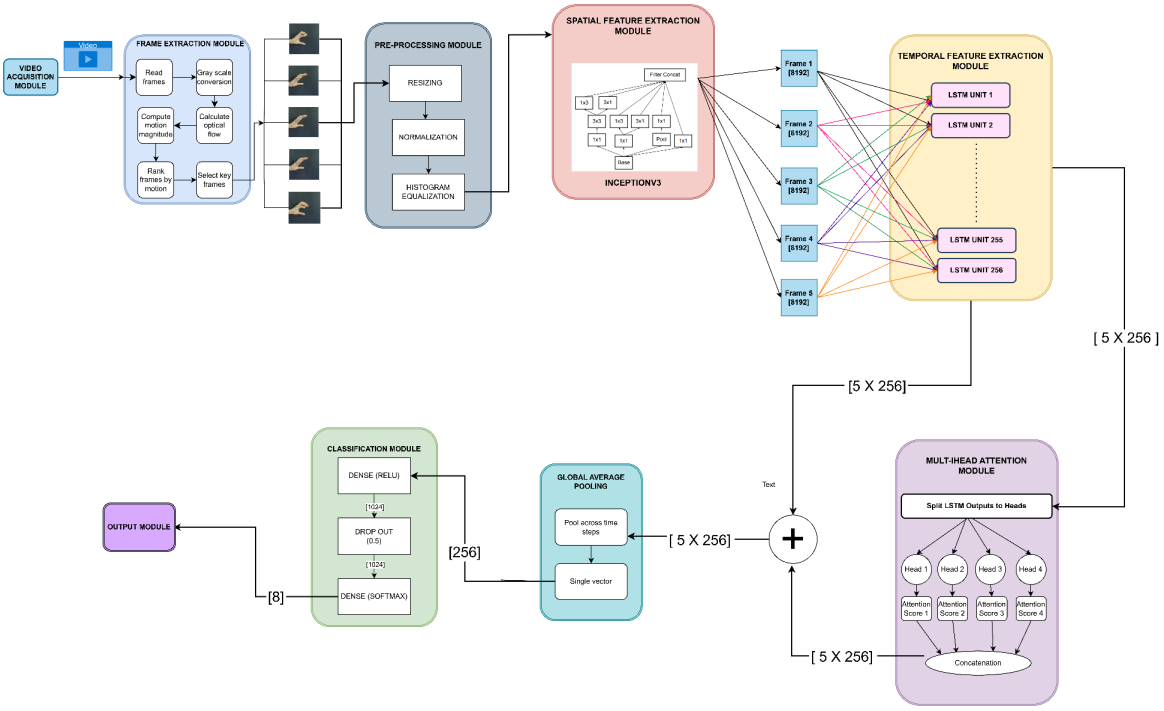


Figure 3: HySTA-Net using InceptionV3

- HySTA-Net using 3D CNN

The proposed architecture integrates a 3D CNN as the primary feature extractor, designed to capture both spatial details and temporal dynamics from input video sequences. The 3D convolutional layers extract meaningful patterns across multiple frames, allowing the model to effectively recognize gesture movements. To enhance

feature stability and ensure consistent scaling, the extracted features are reshaped and passed through a Layer Normalization module. This step ensures that the data maintains a uniform distribution, improving model convergence during training.

The multi-head attention mechanism is incorporated to refine the extracted temporal features. By applying multiple attention heads, the model effectively identifies the most relevant information from the sequence, improving its ability to recognize complex gesture patterns. A Global Temporal Average Pooling layer further compresses these refined features into a compact representation. Finally, the dense layers classify the pooled features, ensuring accurate prediction of sign language gestures. This combination of 3D CNN, attention, and pooling enables the model to achieve strong performance in gesture recognition tasks.

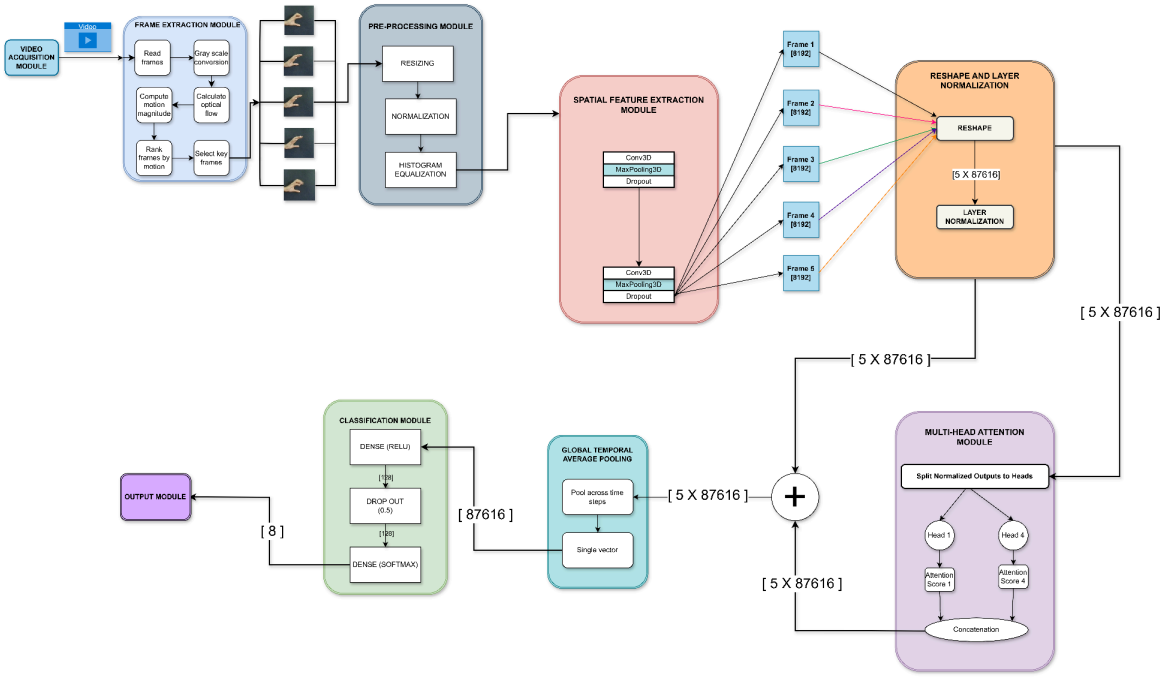


Figure 4 : HySTA-Net using 3D CNN

Chapter 4

Experimental Result Analysis and Discussion

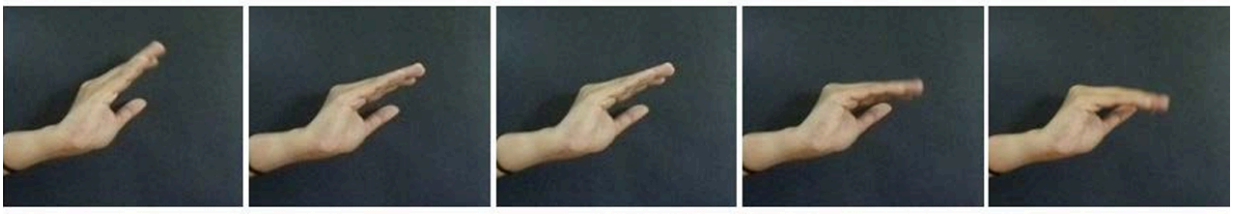
4.1 Dataset

4.1.1 Publicly Available Dataset

This study utilizes a video-based ISL dataset consisting of a total of 412 videos. Researchers focusing on vision-based sign language recognition and hand gesture recognition can benefit significantly from this resource, as it targets the advancement of sign recognition for practical applications. Specifically, the dataset comprises emergency ISL hand gestures, enabling quick and accurate communication in critical scenarios, such as conveying essential messages to authorities.

Eight distinct ISL gestures are included in the dataset: accident, call, doctor, help, hot, lose, pain, and thief. These gestures are commonly used to communicate urgent information or request help in emergency situations. On average, each gesture is represented in about 50 different videos, contributing to the total of 412. The dataset was recorded with 26 adult individuals (12 males and 14 females) aged between 22 and 26 years. Detailed statistics on the number of videos per gesture can be found in Table 1.

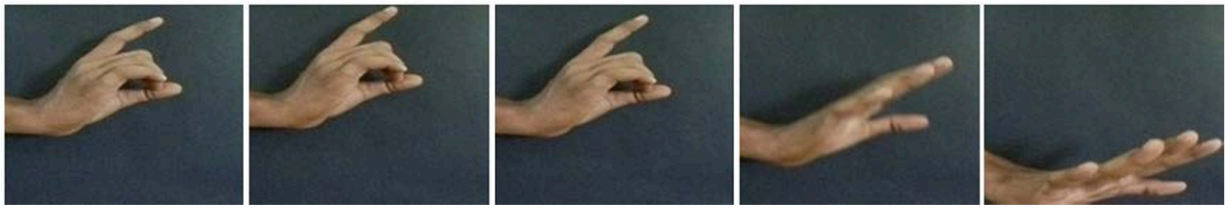
(a) Accident



(b) Call

(c) Doctor

(d) Help



(e) Hot

(f) Lose

(g) Pain

(h) Thief

Figure 5: Eight ISL Emergency Gestures

Table 1 : Distribution of Video Samples Across Different Sign Classes

Class Label	Sign	Number of Videos
0	Accident	52
1	Call	52
2	Doctor	52
3	Help	52
4	Hot	52
5	Lose	50
6	Pain	52
7	Thief	50
	Total	**412**

4.1.2 Custom Dataset

To expand the range of recognizable gestures, a custom ISL dataset comprising 8 different signs was created, with each sign recorded in 50 videos. The dataset captures variations in hand orientation, speed, and lighting conditions, ensuring adaptability to real-world scenarios. Seven signers contributed to the recordings, introducing diversity in hand shapes and movement styles to enhance the model's generalization. Additionally, video resolution and background conditions were standardized to maintain consistency. This curated dataset serves as a crucial resource for assessing the performance of the proposed sign language recognition model.

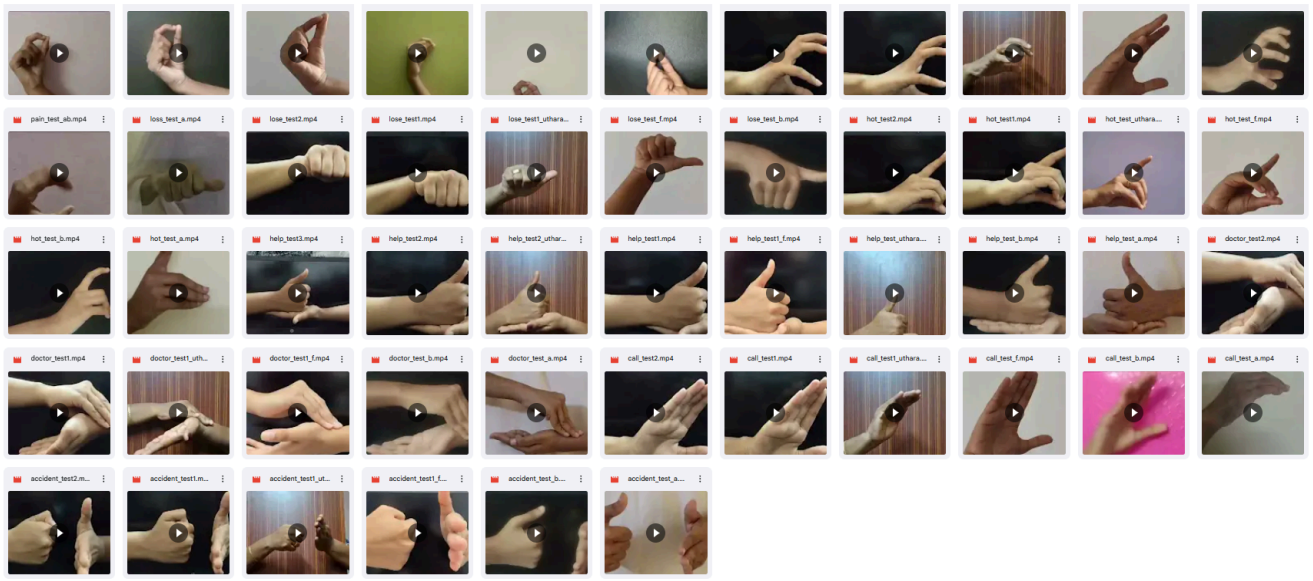


Figure 6: Custom dataset videos

4.2 Key Frame Extraction Analysis

Maintaining important motion cues in dynamic gesture recognition while reducing redundancy requires precise key frame extraction. Several key frame extraction methods were evaluated, including the High Entropy Method, Structural Similarity Index (SSIM), Gradient-Based Method, and Optical Flow Method. Each approach aimed to capture significant gesture transitions while minimising redundancy.

Through experimentation, the Optical Flow Method was identified as the most effective, as it successfully isolated meaningful gesture transitions while reducing redundant frames. It consistently extracted frames that captured significant gesture changes, resulting in higher recognition performance compared to other methods. This improvement was reflected in the overall accuracy and generalization capability of our model. The superior performance of the optical flow method highlights the importance of capturing dynamic information in temporal tasks, making it the optimal choice for our sign language recognition system.

Table 2: Comparison of Key Frame Extraction methods

Key Frame Extraction Method	Custom test accuracy	Model Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Frame Selection	46	82.45	83.12	81.20	82.15
High Entropy Method	73	89.32	90.15	88.75	89.44
SSIM (Structural Similarity)	76	92.18	93.05	91.50	92.27
Gradient-Based Method	82	95.1	95.68	94.85	95.26
Optical Flow Method	**96**	**100**	**100**	**100**	**100**

4.3 Preprocessing Techniques Analysis

The effectiveness of the preprocessing techniques applied during the initial processing stage was analyzed by evaluating their impact on input frame quality. Resizing to 150×150 pixels ensured uniform input dimensions, while normalization improved model convergence by scaling pixel values. Histogram equalization enhanced contrast, making subtle gesture details more distinguishable for better recognition.

Additional preprocessing methods such as Gaussian blur, noise reduction, and data augmentation were not incorporated, as preserving the natural appearance of gestures was essential to maintaining critical features for accurate recognition. The preprocessing techniques applied to the input frames are illustrated in Figure 8 .

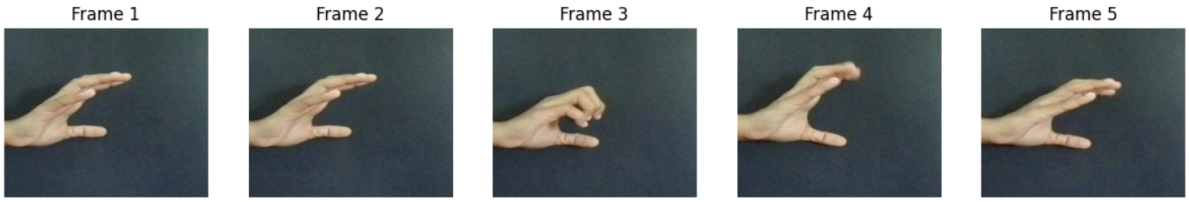


Figure 7 : Input Frames of a sign gesture video

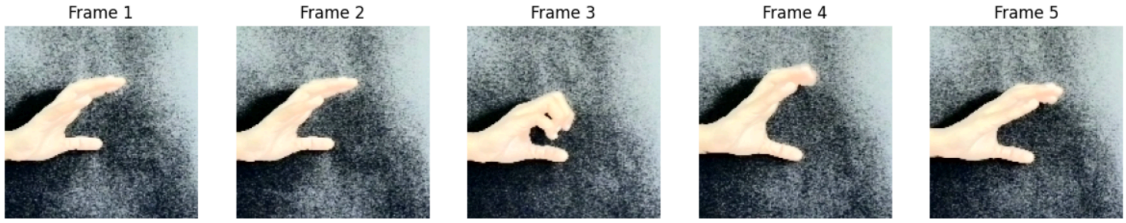


Figure 8: Output Frames of a sign gesture video after resizing, normalization and histogram equalisation

4.4 Performance Evaluation of HySTA-Net

In the study, multiple deep learning models were evaluated to determine their effectiveness in recognizing ISL gestures. The performance of each model was assessed using key metrics such as Precision, Recall, F1-Score, Test Accuracy, and Custom Test Accuracy on both the existing ISL dataset and the custom dataset of 50 signs. This section provides a detailed comparison of model performance, highlighting the strengths and limitations of each approach.

Table 3 presents the performance of baseline models previously reported in existing literature. These models were trained and tested on the original ISL dataset, consisting of 412 videos representing 8 emergency-related gestures. These results serve as a benchmark for comparing the effectiveness of custom models developed in this study.

Table 3: Performance of Baseline Models on the ISL Emergency Gesture Dataset

Model	Precision (%)	Recall (%)	F1-Score (%)	Test Accuracy (%)
VGG16 + LSTM (Existing paper, 2022)	97.5	97.75	97.38	98
3D CNN (Existing paper, 2022)	85.0	83.63	82.63	82

The following tables showcase the performance of various deep learning models developed as part of this research on the custom ISL dataset, which includes 50 unique sign videos captured from 7 signers. The results are divided into two categories to highlight the impact of incorporating attention mechanisms on model performance.

I. Models without Attention Mechanism

The Table 5 presents the results of models that rely solely on convolutional and LSTM layers without any attention mechanism. These models provide a baseline for understanding how much improvement is achieved by adding attention modules. The absence of attention resulted in lower performance on the custom dataset, particularly in complex gestures requiring temporal focus.

Table 4 : Performance Evaluation of Models without Attention Mechanism

Model	Precision (%)	Recall (%)	F1-Score (%)	Test Accuracy (%)	Custom Test Accuracy (50 videos) (%)
HySTA-Net using VGG16	**99**	**98**	**98.8**	**98.8**	**88**

HySTA-Net using ResNet50	96.90	96.39	96.38	96.39	65.79
HySTA-Net using InceptionV3	96	95	95	95	26
HySTA-Net using 3D CNN	86	84	84.34	84.34	18.37

II. Models with Attention Mechanism

The Table 4 focuses on models where an attention mechanism (single-head or multi-head) was integrated with the base architecture. Attention mechanisms are designed to enhance the model's ability to focus on crucial temporal features, leading to improved recognition performance. The inclusion of attention consistently resulted in better or comparable performance, especially in scenarios where gesture variations were subtle.

Table 5: Performance Evaluation of Models with Attention Mechanism

Model	Precision (%)	Recall (%)	F1-Score (%)	Test Accuracy (%)	Custom Test Accuracy (50 videos) (%)
HySTA-Net using VGG16(Multi-Head Attention)	**100**	**100**	**100**	**100**	**96**
HySTA-Net using VGG16(Single-Head Attention)	100	100	100	100	88
HySTA-Net using ResNet50 (Single-Head Attention)	97.93	97.59	97.57	97.59	55.26

HySTA-Net using ResNet50 (Single-Head Attention)	97.93	97.59	97.57	97.59	55.26
HySTA-Net using 3D CNN (Multi-Head Attention)	82	83	82	83.14	22.45

For HySTA-Net using VGG16 (Multi-Head Attention), a confusion matrix was generated (see Figure 8) to provide a detailed view of the classification performance across the 8 gesture classes. The confusion matrix exhibits strong diagonal dominance, indicating that most gestures were correctly classified with very few misclassifications. This robust performance underscores the model's ability to effectively differentiate among similar dynamic gestures.

The findings reveal that VGG16 + LSTM outperformed the 3D CNN model, achieving 98% test accuracy compared to 82%, which highlights superior temporal feature handling. Among the evaluated models, architectures incorporating multi-head attention demonstrated the best performance, with HySTA-Net using VGG16 (Multi-Head Attention) achieving 100% test accuracy and 96% custom dataset accuracy. In contrast, models without attention, despite high test accuracy, showed lower custom test accuracy, suggesting potential overfitting issues. Moreover, the 3D CNN model recorded the weakest performance, with a custom dataset accuracy of only 17%, reflecting its limitations in capturing temporal dependencies effectively. These results emphasize that attention mechanisms significantly enhance the extraction of temporal features and overall model generalization. Additionally, the data indicates that increasing model complexity (as seen with ResNet50) does not necessarily translate to improved performance without effective temporal processing and data diversity.

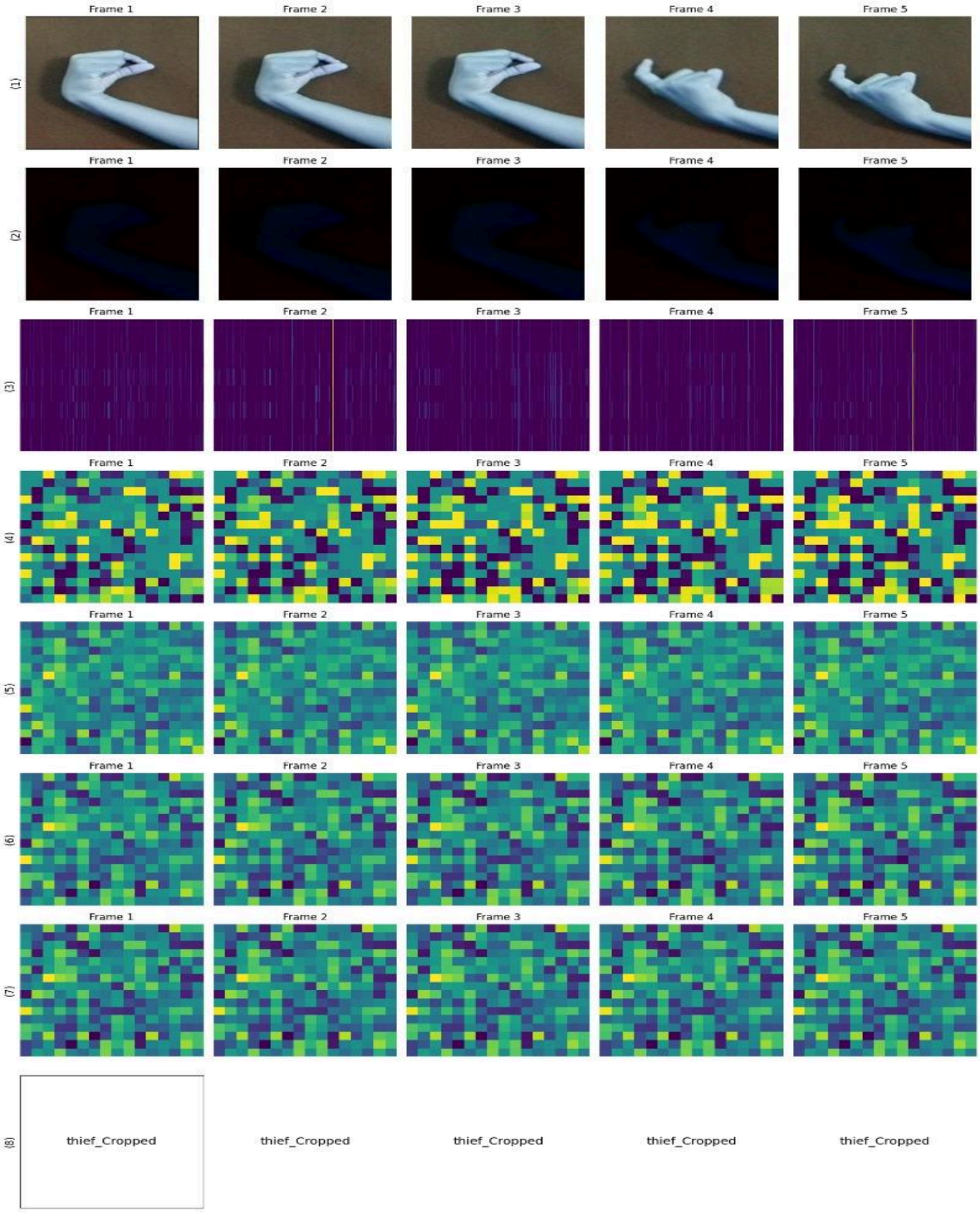


Figure 9 : Sequential outputs from different stages of the proposed ISL gesture recognition pipeline, including (1) Frame extraction, (2) Pre-processing, (3) Feature extraction using VGG16, (4) Temporal modeling with LSTM, (5) Attention mechanism via Multi-Head Attention, (6) Residual connection enhancement, and (7) Final feature aggregation through Global Average Pooling.

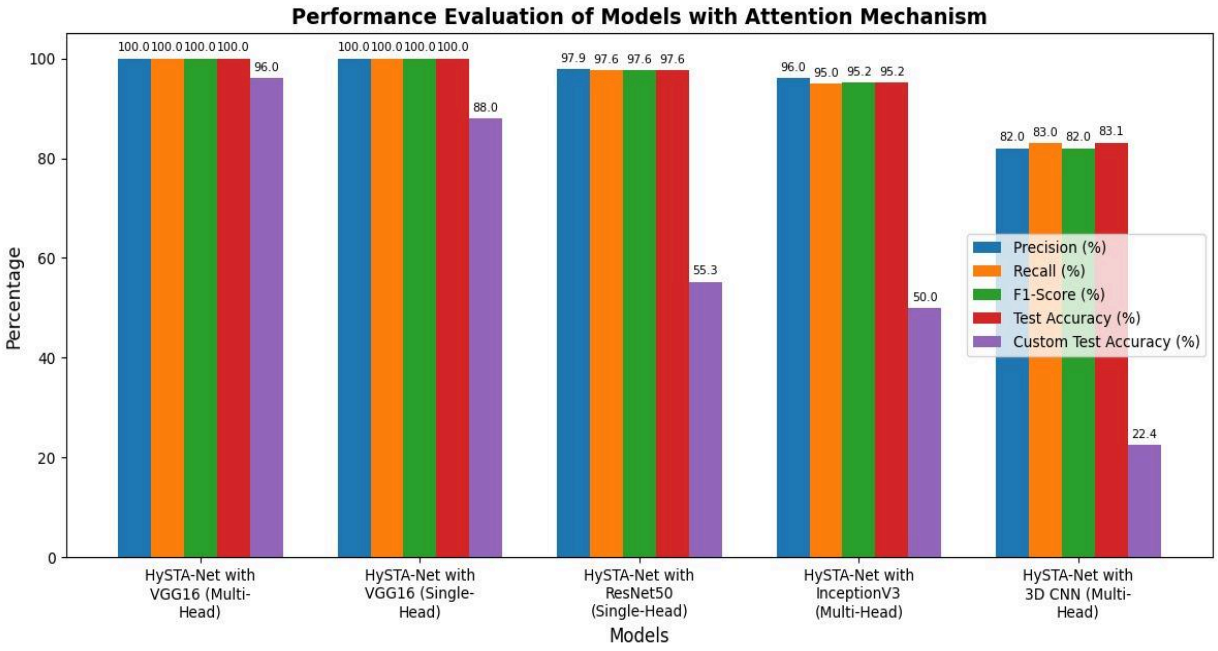


Figure 10 : Performance Evaluation of Models with Attention Mechanism

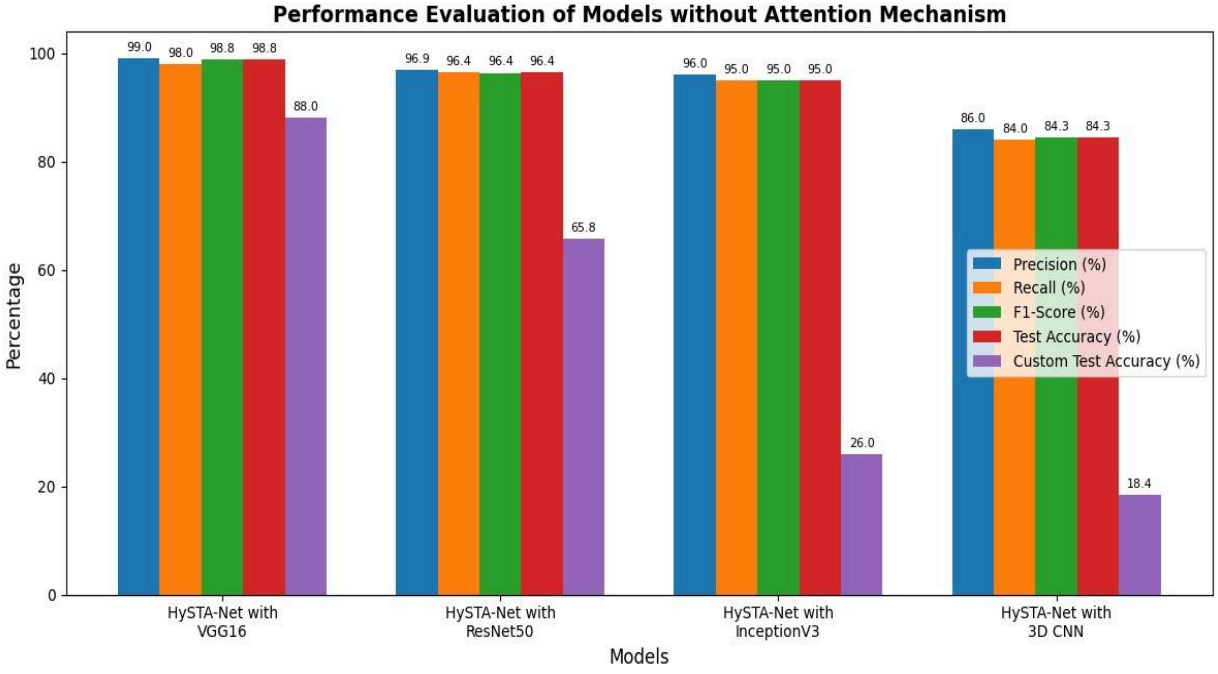


Figure 11 : Performance Evaluation of Models without Attention Mechanism

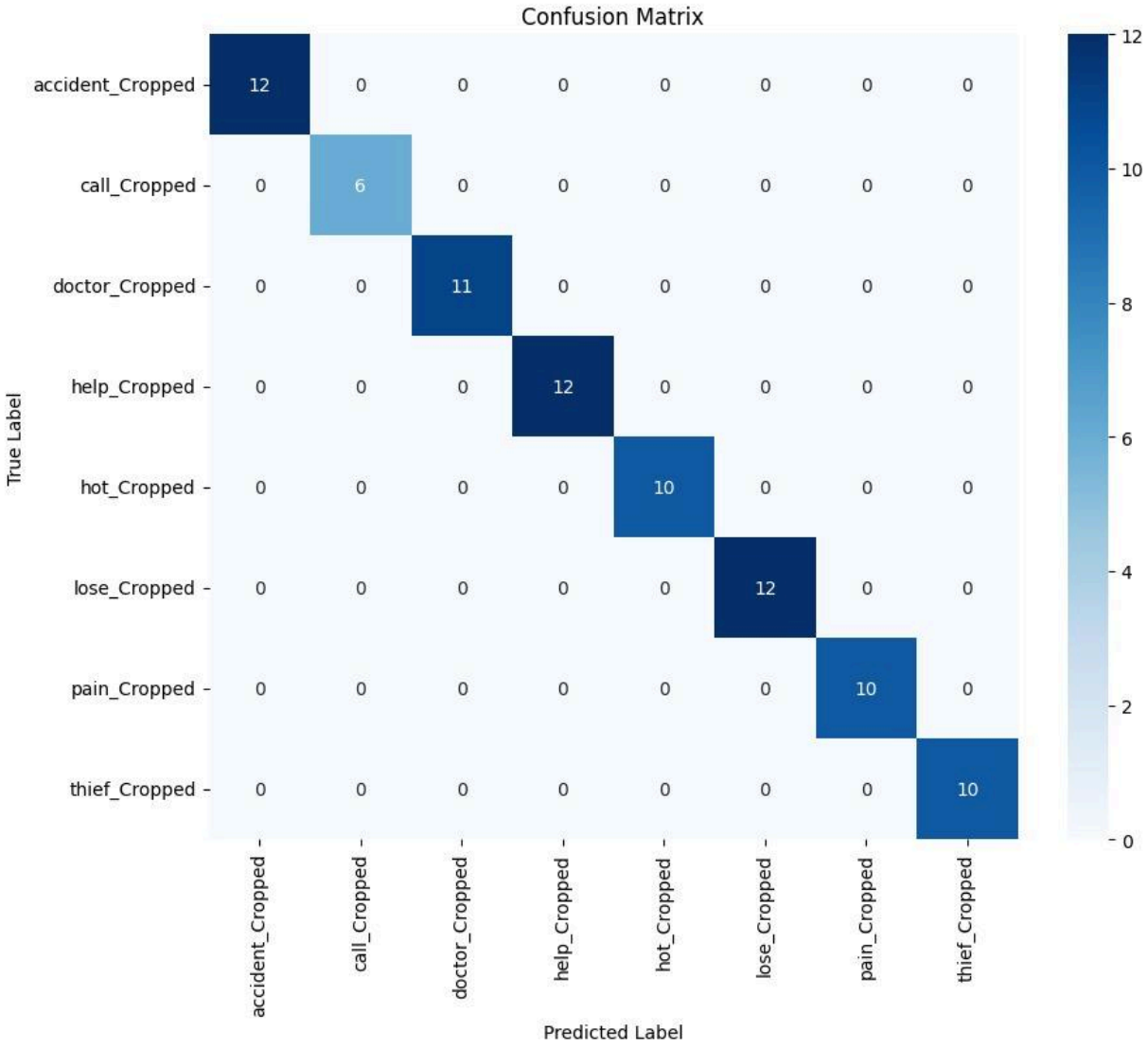


Figure 12 : Confusion Matrix for HySTA-Net using VGG16 (Multi-Head Attention) illustrating class-wise classification performance on the custom ISL dataset for 8 emergency gestures.

Overall, the findings indicate that HySTA-Net using VGG16 with LSTM and Multi-Head Attention provides the best balance between performance and generalization across both datasets, making it the most effective model in the study.

Chapter 5

Conclusion and Future Scope

The recognition of ISL plays a crucial role in bridging the communication gap for the hearing-impaired, particularly in critical emergency scenarios where conveying essential messages quickly is vital. This work aimed to develop an efficient ISL recognition system by experimenting with various deep learning architectures, including VGG16, ResNet50, 3D CNN, and InceptionV4, integrated with temporal models such as LSTM and enhanced by multi-head attention mechanisms. The goal was to accurately classify emergency signs such as accident, help, doctor, and call, which are crucial in real-life urgent situations.

Among the models evaluated, the VGG16 + LSTM + Multi-Head Attention architecture demonstrated superior performance, achieving 100% test accuracy and 96% accuracy on the custom dataset. The integration of multi-head attention significantly improved the capture of complex spatial and temporal features, resulting in better generalization compared to other architectures. Models without attention, despite high test accuracy, exhibited lower performance on the custom dataset, indicating potential overfitting. The 3D CNN model showed the weakest performance, emphasising the limitations of end-to-end volumetric approaches in handling fine-grained temporal dynamics.

Keyframe extraction was another critical component of this project. Several algorithms, including high entropy, SSIM, and optical flow, were explored to identify the most informative frames. The optical flow method outperformed the others by capturing significant motion between frames, contributing to better overall model performance.

This research underscores the importance of combining effective temporal modelling with attention mechanisms to enhance recognition performance in real-world conditions. The findings hold societal relevance by providing a robust platform for communication in emergencies, where rapid and accurate sign recognition can save lives. The system offers

potential for real-world deployment, with scope for improvements by expanding datasets and enhancing adaptability in various environments.

By achieving high accuracy in recognising essential emergency signs, this work takes a step forward in empowering the hearing-impaired community and highlights the transformative impact of deep learning in addressing societal challenges.

To advance this system, future work should focus on expanding the dataset to include a wider range of emergency-related signs, regional ISL variations, and more complex gesture sequences. Real-time implementation on edge devices, such as smartphones and wearables, can enhance accessibility, while multimodal integration using depth cameras and sensor-based gloves may improve recognition accuracy. Exploring Transformer-based models like Vision Transformers and SignBERT could further refine spatial-temporal feature extraction. Additionally, integrating voice-based output and emergency response automation would increase the system's practical utility.

By incorporating these advancements, the ISL recognition system can evolve into a real-time, multimodal tool that significantly improves accessibility and emergency communication for the deaf community.

Chapter 6

References

[1] G. A. Rao, P. V. V. Kishore, A. S. C. S. Sastry, D. A. Kumar, and K. Kumar, ‘‘Selfie continuous sign language recognition with neural network classifier,’’ in Proceedings of 2nd International Conference on Micro-Electronics, Electromagnetics and Telecommunications. Singapore: Springer, 2018, pp. 31–40

[2] N.-T. Do, S.-H. Kim, H.-J. Yang, and G.-S. Lee, “Robust hand shape features for dynamic hand gesture recognition using multi-level feature LSTM”, Appl. Sci., vol. 10, no. 18, p. 6293, Sep. 2020.

[3] R. Cui, H. Liu, and C. Zhang, Recurrent convolutional neural networks for continuous sign language recognition by staged optimization, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 73617369.

[4] A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, “MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences”, Expert Syst. Appl., vol. 139, Jan. 2020, Art. no. 112829.

[5] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, Dynamic sign language recognition based on video sequence with BLSTM-3Dresidualnetworks, IEEE Access, vol. 7, pp. 3804438054, 2019, doi: 10.1109/ACCESS. 2019.2904749.

[6] V. John, A. Boyali, S. Mita, M. Imanishi, and N. Sanma, Deep learning based fast hand gesture recognition using representative frames, in Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA), Nov. 2016, pp. 18.

[7] K. Lai and S. N. Yanushkevich, CNN RNN depth and skeleton based dynamic hand gesture recognition, in Proc. 24th Int. Conf. Pattern Recog nit. (ICPR), Aug.2018,pp. 34513456,doi:10.1109/ICPR.2018.8545718.

[8] F. Obaid, A. Babadi, and A. Yoosofan, Hand gesture recognition in video sequences using

deep convolutional and recurrent neural networks, Appl. Comput. Syst., vol. 25, no. 1, pp. 5761, May 2020, doi: 10.2478/ acss-2020-0007.

[9] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 42074215.

[10] S. Saqib, A. Ditta, M. A. Khan, S. A. R. Kazmi, and H. Alquhayz, ‘‘Intelligent dynamic gesture recognition using CNN empowered by edit distance,’’ Comput., Mater. Continua, vol. 66, no. 2, pp. 2061–2076, 2021.

[11] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, ‘‘Using convolutional 3D neural networks for user-independent continuous gesture recognition,’’ in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 49–54

[12] A. Mujahid, M. J. Awan, A. Yasin, M. A. Mohammed, R. Damaševičius, R. Maskeliunas, and K. H. Abdulkareem, ‘‘Real-time hand gesture recognition based on deep learning YOLOv3 model,’’ Appl. Sci., vol. 11, no. 9, p. 4164, May 2021

[13] M. Khari, A. K. Garg, R. Gonzalez-Crespo, and E. Verdú, ‘‘Gesture recognition of RGB and RGB-D static images using convolutional neural networks,’’ Int. J. Interact. Multimedia Artif. Intell., vol. 5, no. 7, p. 22, 2019

[14] S. Masood, A. Srivastava, H. Thuwal, and M. Ahmad, ‘‘Real-time sign language gesture (word) recognition from video sequences using CNN and RNN,’’ in Intelligent Engineering Informatics, vol. 695, V. Bhateja, C. C. Coello, S. Satapathy, and P. Pattnaik, Eds. Singapore: Springer, 2018, pp. 623–632

[15] V. Adithya and R. Rajesh, ‘‘Hand gestures for emergency situations: A video dataset based on words from Indian sign language,’’ Data Brief, vol. 31, Aug. 2020.

[16] S.Escalera, V. Athitsos, and I. Guyon, Challenges in multi-modal gesture recognition, J. Mach. Learn. Res., vol. 17, no. 2, pp. 154, 2016.

[17] B. Garcia and S. A. Viesca, Real-time American sign language recognition with convolutional neural networks, Convolutional Neural Netw. Vis. Recognit., vol. 2, pp. 225232, 2016.

[18] C. A. Padden, Sign language geography, in Deaf Around the World: The Impact of Language. New York, NY, USA: Oxford Univ. Press, 2011, pp. 1937.

[19] D. Tirthankar, S. Sambit, K. Sandeep, D. Synny, and B. A. Anupam, "Multilingual multimedia Indian sign language dictionary tool", in Proc. 6th Workshop Asian Lang. Resour., 2008, pp. 1112.

[20] Y. Fang, K. Wang, J. Cheng, and H. Lu, "A real-time hand gesture recognition method", in Proc. IEEE Int. Conf. Multimedia Expo, Beijing, China, Jul. 2007, pp. 995998.

[21] Oudah, A. Al-Naji, and J. Chahl, Hand gesture recognition based on computer vision: A review of techniques, J. Imag., vol. 6, no. 8, p. 73, 2020.

[22] H. Grant and C.-K. Lai, Simulation modeling with artificial reality technology (SMART): An integration of virtual reality and simulation modeling, in Proc. Winter Simulation Conf., Washington, DC, USA, Dec. 1998, pp. 437441.

[23] T.-D. Tan and Z.-M. Guo, Research of hand positioning and gesture recognition basedonbinocularvision, inProc.IEEEInt.Symp.VRInnov., Singapore, Mar. 2011, pp. 311315.

[24] S.-H. Lee, M.-K. Sohn, D.-J. Kim, B. Kim, and H. Kim, Smart TV interaction system using face and hand gesture recognition, in Proc. IEEE Int. Conf. Consum. Electron. (ICCE), Las Vegas, NV, USA, Jan. 2013, pp. 173174

[25] J. Huang, W.Zhou,H.Li,andW.Li, Sign Language recognition using 3D convolutional neural networks, in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Turin, Italy, Jun. 2015, pp. 16

[26] R. Cui, H. Liu, and C. Zhang, Recurrent convolutional neural networks for continuous sign language recognition by staged optimization, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 73617369.

[27] H. Liu and L. Wang, Gesture recognition for human-robot collaboration: Areview, Int. J. Ind. Ergonom., vol. 68, pp. 355367, Nov. 2018.

[28] D.Xu,X.Wu,Y.-L.Chen,andY.Xu, Online Dynamic Gesture Recognition for human robot interaction, J. Intell. Robot. Syst., vol. 77, nos. 34, pp. 583596, Mar. 2015.

[29] N. L. Hakim, S.-W. Sun, M.-H. Hsu, T. K. Shih, and S.-J. Wu, Virtual guitar: Using real-time finger tracking for musical instruments, Int. J. Comput. Sci. Eng., vol. 18, no. 4, pp. 438450, 2019.

[30] N.DawarandN.Kehtarnavaz, Real-time continuous detection and recognition of subject-specific smart TV gestures via fusion of depth and inertial sensing, IEEE Access, vol. 6, pp. 70197028, 2018.

[31] W. Kaczmarek, J. Panasiuk, S. Borys, and P. Banach, Industrial robot control by means of gestures and voice commands in off-line and on-line mode, Sensors, vol. 20, no. 21, p. 6358, Nov. 2020.

[32] P. Neto, M. Simão, N. Mendes, and M. Safeea, Gesture-based human robot interaction for human assistance in manufacturing, Int. J. Adv. Manuf. Technol., vol. 101, nos. 14, pp. 119135, Mar. 2019.

[33] Ç. Gökçe, O. Özdemir, A.A. Kındıroğlu, L. Akarun, Score-level multi cue fusion for sign language recognition, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 294–309.

[34] K. Papadimitriou, G. Potamianos, Multimodal sign language recognition via temporal deformable convolutional sequence learning, in: Interspeech, 2020, pp. 2752–2756.

[35] S. Zhang, W. Meng, H. Li, X. Cui, Multimodal spatiotemporal networks for sign language recognition, IEEE Access 7 (2019) 180270–180280.

[36] S. Ravi, M. Suman, P. Kishore, K. Kumar, A. Kumar, et al., Multi modal spatio-temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition, J. Comput. Lang. 52 (2019) 88–102.

[37] Y. Liao, P. Xiong, W. Min, W. Min, J. Lu, Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks, IEEE Access 7 (2019) 38044–38054.

[38] N.C. Camgoz, O. Koller, S. Hadfield, R. Bowden, Sign language transformers: Joint end-to-end sign language recognition and translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10023–10033.

[39] F. Xiao, C. Shen, T. Yuan, S. Chen, CRB-net: A sign language recognition deep learning strategy based on multi-modal fusion with attention mechanism, in: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2021, pp. 2562–2567.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

Chapter 7

APPENDICES

7.1 Appendix I: Vision, Mission and Program Educational Objectives(PEOs)

VISION

To be a centre of excellence imparting quality education in Computer Science and Engineering and transforming students to critical thinkers and lifelong teams capable of developing environment friendly and economically feasible solutions to real world problems

MISSION

- To provide a strong foundation in Computer Science and Engineering, prepare students for professional career and higher education, and inculcate research interest.
- To be abreast of the technological advances in a rapidly changing world.
- To impart skills to come up with socially acceptable solutions to real world problems, upholding ethical values.

PROGRAMME EDUCATIONAL OBJECTIVES(PEOs)

PEO 1: Excel in professional career by acquiring knowledge in mathematics, science and engineering and applying the knowledge in the design of hardware and software solutions for challenging problems of the society.

PEO 2: Pursue higher studies and research thereby engaging in lifelong learning by adapting to the current trends in the area of Computer Science and Engineering.

PEO 3: Ability to Provide socially acceptable and economically feasible computer oriented solutions to real world problems with teamwork, while maintaining environmental balance, quality and cognizance of the underlying principles of ethics.

7.2 Appendix I: Program Outcomes

1. **PO1- Engineering Knowledge:** Apply mathematical, scientific, and engineering principles, along with specialization in machine learning and computer vision, to create solutions for complex sign language recognition challenges.
2. **PO2- Problem Analysis:** Identify and analyze specific challenges in ISL recognition, such as variability in hand gestures, dynamic nature of signs and the need for real-time processing, drawing substantiated conclusions using foundational knowledge in engineering and data science.
3. **PO3- Design/Development of Solutions:** Designing an ISL recognition system for emergency signs that ensures accurate and real-time recognition while addressing critical communication needs in emergency scenarios with a focus on public safety and accessibility.

4. **PO4- Investigation of Complex Problems:** Use research-based methodologies to design experiments, collect and analyze data, and synthesize results to draw valid conclusions regarding the performance and accuracy of the gesture recognition system.

5. **PO5- Modern Tool Usage:** Utilize advanced deep learning frameworks and computer vision tools, including neural network architectures and optical flow techniques, for the development and real-time testing of the ISL emergency sign recognition system, while considering their limitations.

6. **PO6- The Engineer and Society:** Apply knowledge of social and cultural contexts to ensure that the ISL emergency sign recognition system meets the diverse needs of the hearing-impaired community, promoting accessibility and inclusivity.

7. **PO7- Environment and Sustainability:** Understand the societal and environmental implications of technology in communication accessibility and strive for sustainable solutions by developing a resource-efficient and scalable system.

8. **PO8- Ethics:** Commit to ethical principles in the design and implementation of the ISL Emergency Sign Recognition system, ensuring user data protection and privacy.

9. **PO9- Individual and Team Work:** Operate effectively both as an individual and as part of a multidisciplinary team, contributing knowledge in machine learning, computer vision, and user-centric design for the development of the ISL Emergency Sign Recognition System.

10. **PO10- Communication:** Communicate complex engineering concepts, system functionalities, and outcomes effectively to both technical and non-technical stakeholders, fostering understanding and support for the ISL recognition project.

11. **PO11- Project Management and Finance:** Apply principles of engineering management to organize and implement the project efficiently, working within budgetary constraints to create a functional and effective gesture recognition system.

12. **PO12- Life-long Learning:** Recognize and prepare for the need to continuously learn and adapt to emerging tools and technologies in machine learning, gesture recognition,

and human-computer interaction to improve and expand the capabilities of the ISL Emergency Sign Recognition System.

7.3 Appendix II: Course Outcomes (COs)

1. **CO1**- Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: **Apply**).
2. **CO2-** Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: **Apply**).
3. **CO3**- Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: **Apply**).
4. **CO4**- Plan and execute tasks utilising available resources within timelines, following ethical and professional norms (Cognitive knowledge level: **Apply**).
5. **CO5**- Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: **Analyze**).
6. **CO6-** Organise and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: **Apply**).

7.4 Appendix III: Fulfilment of Programme Outcomes (COs)

1. PO1 - Engineering Knowledge

The work applies mathematical, scientific, and engineering principles to develop a robust Indian Sign Language (ISL) emergency sign recognition system. By utilizing deep learning architectures such as VGG16, ResNet50, InceptionV3, and 3D CNN integrated with LSTM and

multi-head attention, the system effectively captures spatial and temporal features for accurate gesture classification. Optical flow-based keyframe extraction enhances motion representation, reducing redundancy while preserving essential movement cues. This work demonstrates the integration of computer vision and artificial intelligence to address real-world challenges, improving accessibility and communication for the hearing-impaired in emergency situations.

2. PO2 - Problem Analysis

The work addresses key challenges in dynamic ISL gesture recognition, such as variations in hand movements, background complexity, and lighting conditions, which can impact model accuracy. To overcome these challenges, preprocessing techniques like resizing, normalization, and contrast enhancement are applied to improve data consistency. By directly extracting spatial-temporal features using deep learning models such as VGG16 + LSTM with Multi-Head Attention, the system effectively captures gesture dynamics while minimizing the impact of environmental factors. This comprehensive approach enhances the robustness and generalizability of the ISL Emergency Sign Recognition System.

3. PO3 - Design and Development of Solutions

The solution is systematically developed in distinct phases: dataset preprocessing, model training, and evaluation. A publicly available ISL dataset is used for training, while a custom dataset is captured to evaluate model performance across different backgrounds, lighting conditions, and signers. Key frames are extracted from videos using an optimized optical flow-based method, followed by preprocessing techniques such as resizing, normalization, and histogram equalization to enhance input quality. Deep learning architectures, including VGG16 + LSTM with Multi-Head Attention, are trained to effectively capture spatial and temporal dependencies. This structured approach ensures a robust and generalizable ISL recognition system capable of accurately identifying emergency signs in diverse real-world conditions.

4. PO4 - Conduct Investigations of Complex Problems

The work addresses key challenges in dynamic Indian Sign Language (ISL) gesture recognition, focusing on selecting the most informative frames and enhancing temporal

modeling for improved classification accuracy. Deep learning architectures, including VGG16 + LSTM + Multihead Attention, InceptionV3 + LSTM + Multihead Attention, 3D CNN + Multihead Attention, and ResNet50 + LSTM + Self-Attention, are evaluated to determine their effectiveness in recognizing emergency ISL gestures. The models are trained on a publicly available dataset and tested on a custom dataset featuring variations in background, lighting conditions, and different signers to assess their generalization capabilities. Performance metrics such as accuracy, precision, and recall are analyzed to optimize the model, ensuring a robust and adaptable recognition system for real-world emergency scenarios.

5. PO5 - Modern Tool Usage

The work utilizes advanced deep learning libraries and tools, including TensorFlow and Keras for model training, OpenCV for video preprocessing, and NumPy for efficient data manipulation. Keyframe extraction techniques such as optical flow and structural similarity (SSIM) are employed to enhance feature selection, ensuring that the most relevant frames contribute to classification. The deep learning models, VGG16 + LSTM + Multihead Attention, InceptionV3 + LSTM + Multihead Attention, 3D CNN + Multihead Attention, and ResNet50 + LSTM + Self-Attention are implemented to capture spatial-temporal dependencies in dynamic ISL emergency gestures. This integration of modern tools enables effective gesture recognition, improving model accuracy and generalizability across diverse conditions.

6. PO6 - The Engineer and Society

The work addresses a critical societal need by developing a system that enhances communication for the deaf and hard-of-hearing community through dynamic ISL Emergency gesture recognition. By focusing on emergency gestures, the system promotes accessibility and inclusivity, ensuring effective interaction in critical situations. Its potential applications extend to emergency response, public services, and assistive technologies, reinforcing its societal impact. This initiative aligns with the broader goal of leveraging engineering and artificial intelligence to create meaningful, real-world solutions that serve the community.

7. PO7 - Environment and Sustainability

The work prioritizes environmental sustainability by relying on software-based solutions for gesture recognition, thereby avoiding additional hardware requirements. Since the system is implemented on standard devices such as laptops and smartphones, it minimizes resource consumption and environmental impact. The emphasis on a software-only solution aligns with sustainable engineering practices, ensuring that the technology remains accessible and environmentally responsible.

8. PO8 - Ethics

The work upholds high ethical standards by ensuring that all data used in training is collected ethically and used solely for research and development purposes. Intellectual property rights are respected through proper citations of existing work in sign language recognition. Throughout the project, data privacy and responsible use of information are maintained, demonstrating the team's commitment to ethical practices, transparency, and integrity in the research process.

9. PO9 - Individual and Team Work

The work exemplifies effective teamwork and collaboration, with each team member contributing to critical aspects such as dataset creation, model training, and evaluation. Tasks were divided strategically to ensure a cohesive workflow, and regular meetings facilitated knowledge sharing and progress updates. This collective effort fosters a cohesive and collaborative approach towards achieving the objectives of the study.

10. PO10 - Communication

The research study effectively communicates complex engineering activities to the engineering community and society at large. In addition to the abstract, the researchers likely communicate their work through technical reports, research papers and presentations. The study underwent four internal evaluations conducted by an evaluation committee, indicating the emphasis placed on effective communication within the research community.

11. PO11 - Project management and finance

The research study demonstrates effective project management practices and considerations. The researchers meticulously plan and execute various phases of the study, including input preprocessing, frame selection, region selection, and data embedding and allocate resources, such as cover videos, input datasets, and computational resources, to conduct experiments and evaluate the proposed methodology. By employing systematic project management techniques, the researchers ensure the efficient and organised progression of the study, facilitating successful outcomes and contributing to effective multidisciplinary collaboration.

12. PO12 - Life-long learning

The research study demonstrates the researchers' engagement in independent and lifelong learning. The study involves a thorough review of existing literature to establish the current state of video steganography. The researchers identify limitations and gaps in existing methods and propose an innovative methodology to address them. Through their research efforts, the study contributes to the broader context of technological change and emphasises the importance of continuous learning and adaptation.