Deep Spatiotemporal Network Based Pakistan Hand Sign Language Recognition

Author Shehryar Naeem

BSc (Hons.) Computer Science
Deliverable 1: Final Year Dissertation

Supervised by Dr. Md Azher Uddin

Heriot-Watt University
School of Mathematical and Computer Sciences

November 2024

DECLARATION

I, Shehryar Naeem, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed:

Date: 21/11/2024

ABSTRACT

Sign language is the key means of communication within the deaf and hard-of-hearing communities. Despite this value, Pakistan Sign Language (PSL) remains relatively unexplored in the sphere of sign language recognition due to the lack of a standardized dataset and suitable models with the ability to extract temporal motion information. Motivated by this, a novel video-based deep spatiotemporal framework is proposed for improving the verification of PSL. The proposed approach will be developed on the integration of ResNet-101, extracting spatial features, and Motion Binary Pattern (MBP) for capturing temporal dynamics. Features extracted are sent through classification models in sequence, including Transformers, Bidirectional Long Short-Term Memory (Bi-LSTM), and Gated Recurrent Unit (GRU). Final prediction is done based on the majority voting mechanism for improved reliability. This has been tested rigorously on the PkSLMNM dataset, comprising a video-based PSL dataset with much sign variation from multiple participants. The proposed system provides state-of-the-art methodologies toward addressing a few challenges in PSL recognition to ensure inclusivity and improve communication among the deaf, while offering a framework for future technological advancement.

Keywords: Pakistan Sign Language (PSL), PSLR, Pakistan Sign Language Recognition, Spatial-Temporal Network, Feature Extraction, Bidirectional Long Short-Term Memory (Bi-LSTM), Gesture Recognition, ResNet-101, Motion Binary Pattern (MBP), Gated Recurrent Unit (GRU), Transformers

ACKNOWLEDGEMENTS

I want to thank everyone who loves me including my family, my cats and my valuable friends for believing and being there for me. I would also like to take this time to thank my supervisor Dr. Azher. His consistent advice throughout this project has been invaluable and his mentorship has not only guided me academically but also inspired me to approach complex ideas with confidence.

Finally, I would like to thank all professors in our course F20PA and Heriot-Watt for providing me this opportunity to learn and display my abilities.

Table of Contents

List of Figures

List of Tables

ABBREVIATIONS

2D-CNN 2D-Convolutional Neural Networks. 5, 12

ArSL Arabic Sign Language. vii, 3, 8, 12
ASL American Sign Language. vii, 3, 9, 11, 12

Bi-LSTM Bidirectional Long Short-Term Memory. iii, 1, 2, 6, 8, 11, 12, 14, 15, 20, 24
BOVW Bag of Visual Words. 12
BOW Bag of Words. 4, 6

C3D Convolutional 3D. 5, 12
CNN Convolutional Neural Network. 5, 9–12
CPU Central Processing Unit. 16
CSOM Convolutional Self-Organising Map. 8, 12

DCT Discrete Cosine Transform. 8, 12

GCN Graph Convolutional Network. 10, 12
GDPR General Data Protection Regulation. 29
GPU Graphics Processing Unit. 16
GRU Gated Recurrent Unit. iii, ix, 1, 2, 7, 11, 12, 14, 15, 20, 24

HMM Hidden Markov Model. 8, 12
HOG Histogram of Oriented Gradients. 8, 12

I3D Two-Stream Inflated 3D ConvNet. 5, 9, 11, 12
ISL Indian Sign Language. vii, 3, 5, 6, 8, 11, 12

KSL Korean Sign Language. vii, 3, 10, 12

LSTM Long Short-Term Memory. ix, 5, 7–12, 14

MBP Motion Binary Pattern. iii, 1, 2, 11, 13, 15, 20, 24

PADEAF Pakistan Association of the Deaf. 1
PDPA Pakistan's Personal Data Protection Act. 29
PkSLMNM Pakistan Sign Language Manual and Non-Manual. iii, ix, xi, 2, 3, 11–13, 15, 17, 29
PLES Professional, Legal, Ethical, and Social. 2

PSL Pakistan Sign Language. iii, vii, ix, 1–4, 11–14, 16–20, 24, 29
PSLR Pakistan Sign Language Recognition. iii, 1, 20

RAM Random Access Memory. 16
RPN Region Proposal Network. 3, 12

SIFT Scale-Invariant Feature Transform. 6, 12
SURF Speeded-Up Robust Features. 4, 12
SVM Support Vector Machine. 4, 5, 12

TCN Temporal Convolution Networks. 8
TGCN Temporal Graph Convolutional Network. 9, 12
TSM Temporal Shift Module. 5, 12

VGG-19 Visual Geometry Group. 6, 12

WHO World Health Organization. 1

1 INTRODUCTION

Sign language is an essential form of communication for the deaf and hard-of-hearing communities. In 2021, World Health Organization (WHO) declared that around 430 million people suffer from moderate to severe hearing loss [Organization et al. 2021]. Sign language makes use of both manual and non-manual gestures where the former include hands and signs, the latter however uses the upper body and facial expressions [Sameena Javaid 2023]. Several different regions have their own sign languages, each with unique dialects and grammatical nuances [Kaur et al. 2023]. Sign language recognition has become an essential field of study to enhance communication accessibility.

In this field, there are two primary methods: sensor-based and vision-based. In sensor-based method sensors are physically attached to users to record position, motion as well as trajectories of fingers and hand data like seen done via an armband made by Shin et al. [2017]. In vision-based approach there have been research done with static gestures which utilize images [Shah et al. 2023; Singla et al. 2024] and dynamic which uses video data as input. The key distinction between sensor-based and vision-based approaches is how data is gathered and preprocessed [Cheok et al. 2019].

According to Pakistan Association of the Deaf (PADEAF), there are around 250000 individuals in Pakistan who have hearing disabilities and their primary form of communication is Pakistan Sign Language (PSL) [PADEAF [n. d.]]. Not much extensive research is done on Pakistan Sign Language Recognition (PSLR) using both static and dynamic data. We will only be doing a vision-based approach and more specifically using video data and in the next section will be exploring related works done. Current models often fall short in capturing both spatial and temporal features simultaneously, especially for PSL, where there is a lack of standardized datasets and complex sign variations across individuals. The proposed framework includes the use of the combination of pretrained ResNet-101 [He et al. 2016] and Motion Binary Pattern (MBP) [Baumann et al. 2014] as our feature extractors and for classification we will use Transformer [Vaswani et al. 2017], Bidirectional Long Short-Term Memory (Bi-LSTM) [Graves and Schmidhuber 2005] and Gated Recurrent Unit (GRU) [Cho et al. 2014] where a majority voting mechanism will be implemented to enhance prediction reliability.

1.1 Aim and Objectives

The aim of this dissertation is to improve PSLR through a framework that combines spatial and temporal feature extraction methods, sequential classification models and appropriate evaluation metrics. The specific objectives are:

- Develop an efficient end-to-end framework for PSL recognition;
- Use Top K Key frame extraction to select key frames from videos;
- Use ResNet-101 to extract spatial features from key frames;

- Use MBP to capture temporal information;
- Implement Transformer, GRU and Bi-LSTM models and use majority voting to predict PSL signs;
- Train the proposed framework using the PkSLMNM dataset;
- Evaluate the performance of the proposed framework based on empirical classification metrics;
- Optimize the model hyperparameters through rigorous testing on the PkSLMNM dataset;
- Compare the model's accuracy against state-of-the-art models;

1.2 Organisation

This dissertation will follow an organized structure: In Section 2 we will examine the existing research on sign language recognition approaches for video data. Then in Section 3 we describe the proposed approach while Section 4 discusses the project's requirements analysis. Section 5 covers the dataset and evaluation measures. The conclusion is discussed in Section 6. In the back matter Appendix A discusses project management, including risk analysis and Gantt charts for the project timeline. Appendix B discusses the Professional, Legal, Ethical, and Social (PLES) considerations relevant to this research.

2 LITERATURE REVIEW

2.1 Overview

Communication in sign language is both manual and non-manual. As briefly mentioned in Section 1, manual ones include signs which can be formed by the use of hands, whereas non-manuals include head movements, facial expressions, shoulder shrugs, and other forms of body language that add meaning to the context [Sameena Javaid 2023]. These can be captured via two main techniques: vision-based and sensor-based. In this review, we narrow our interest to the vision-based technique more precisely video-based. Researchers from different backgrounds have employed various strategies in overcoming this barrier. Our review will discuss the previous work done on various methods done for various sign languages presenting architectures, datasets used, results found, merits, and demerits of each and every technique.

The review will focus on prior research in the following sequence:

- Pakistan Sign Language (PSL)
- Indian Sign Language (ISL)
- Arabic Sign Language (ArSL)
- American Sign Language (ASL)
- Korean Sign Language (KSL)

Let us take a look at PSL related works first. It is to be noted that each author has taken distinct approaches suitable to the datasets they chose to work upon.

2.2 Pakistan Sign Language (PSL)

Sameena Javaid [2023] introduced a novel framework for PSL recognition using the PkSLMNM dataset, consisting of dynamic video-based gestures from 180 participants. As seen in Figure 1, I3d-ShuffleNet (inspired from [Carreira and Zisserman 2017]) was employed for feature extraction to capture spatiotemporal information from the videos, alongside data augmentation techniques such as flipping, cropping and contrast adjustment. Action Transformer with Region Proposal Network (RPN) was used for gesture classification with an attention mechanism for bounding box generation. The model achieved a testing accuracy of 82.66% and training accuracy of 86.12%. However, the small dataset size and issues like motion blur limit the model's effectiveness in real-world scenarios.

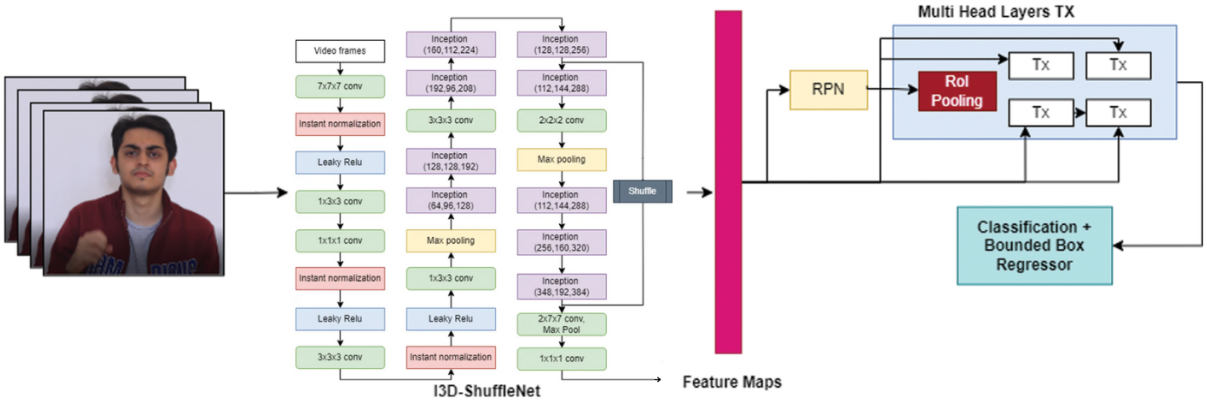

Fig. 1. System architecture of the proposed model by Sameena Javaid [2023]

Mirza et al. [2022] proposed a vision-based system using the self-collected dataset of 5120 static images and 353 dynamic videos of PSL signs from 10 native signers. As seen in Figure 2 the dataset was pre-processed first by resizing, then converting to grayscale images and later performing threshold-based segmentation. Then, features were extracted using Speeded-Up Robust Features (SURF) and clustered by K-means++ to form the BOW model. Classification is later done by the use of a Support Vector Machine (SVM). This in essence, allows the system to achieve an accuracy of 97.80% and 96.53% for static and dynamic signs, respectively.

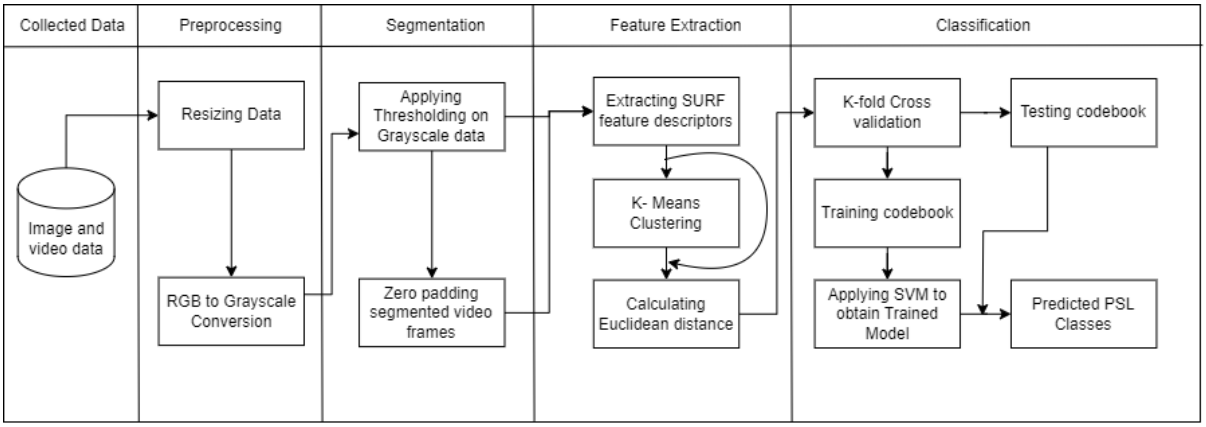

Fig. 2. Proposed PSL recognition flowchart by Mirza et al. [2022]

Hamza and Wali [2023] addressed this challenge by the recognition of PSL on the limited dataset consisting of 80 signed words each having two samples per word. The authors have used data augmentation techniques to help improve model performance by adjusting rightness, rotation, scaling, and translation. They tested three different deep learning models, namely

Convolutional 3D (C3D), Two-Stream Inflated 3D ConvNet (I3D) [Carreira and Zisserman 2017], and a new approach was introduced via the Temporal Shift Module (TSM). Among them, C3D does the best with an accuracy of 93.33%, I3D reaches 87.50%, while TSM performs the poorest with only 35.83% accuracy.

2.3 Indian Sign Language (ISL)

Mittal et al. [2019] introduced a modified Long Short-Term Memory (LSTM) model for continuous sign language recognition using a Leap Motion sensor. The dataset was collected and included 942 sentences of ISL, capturing the 3D coordinates of fingertips. For feature extraction, 2D-Convolutional Neural Networks (2D-CNN) was employed. The modified 4 gated LSTM with 2D-CNN was proposed, featuring 3 layers and a reset gate, which achieved 72.3% accuracy for continuous sentences and 89.5% for isolated words.

Aparna and Geetha [2020] developed a dataset of six isolated words, containing 20 training videos and 10 testing videos for each word. The authors used Inception V3, a pre-trained Convolutional Neural Network (CNN) model for extracting features from video frames, converting video frames to feature vectors and further feeding the output to a stacked LSTM, recognizing temporal features. The model achieved a good accuracy of 94% on the training set.

Adithya and Rajesh [2020] developed a unique video dataset of hand gestures focusing on emergency-related words. The dataset as shown in Figure 3, contains 824 videos of eight hand gestures, such as "help" and "doctor," captured from 26 participants. They proposed two approach for ISL recognition: Traditional feature based and deep learning . For feature extraction, 3D wavelet transform descriptors were used on key frames extracted through image entropy and clustering methods, classified using SVM and the deep learning model investigated using GoogleNet (pretrained CNN) and LSTM , achieving accuracy rates of 90% and 96.25%, respectively.

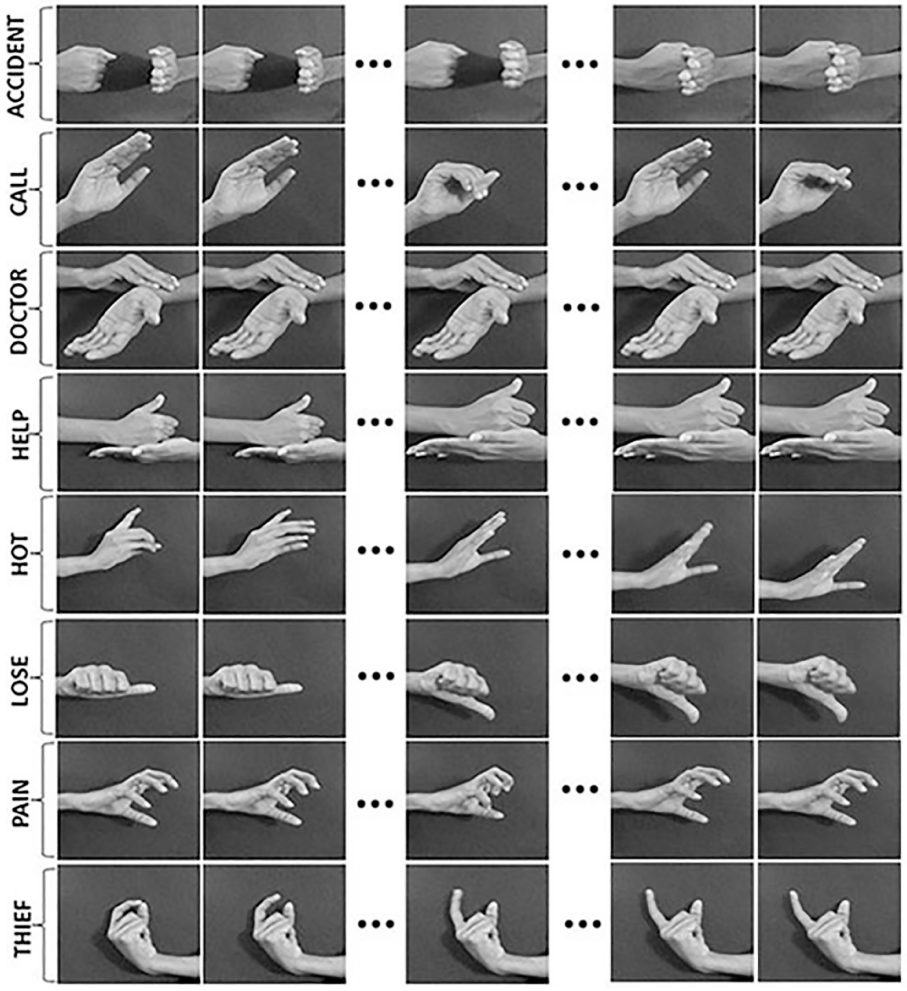

Fig. 3. Dataset created by Adithya and Rajesh [2020]

The work by Das et al. [2023] extends that of Adithya and Rajesh [2020] using the same dataset of 824 videos related to hand gestures dealing with emergencies in ISL, proposing a hybrid approach that fuses Scale-Invariant Feature Transform (SIFT) and BOW together with Visual Geometry Group (VGG-19) for feature extraction. The classification model used was Bidirectional Long Short-Term Memory (Bi-LSTM). That gave an average accuracy of 94.42% using a Bi-LSTM network for classification.

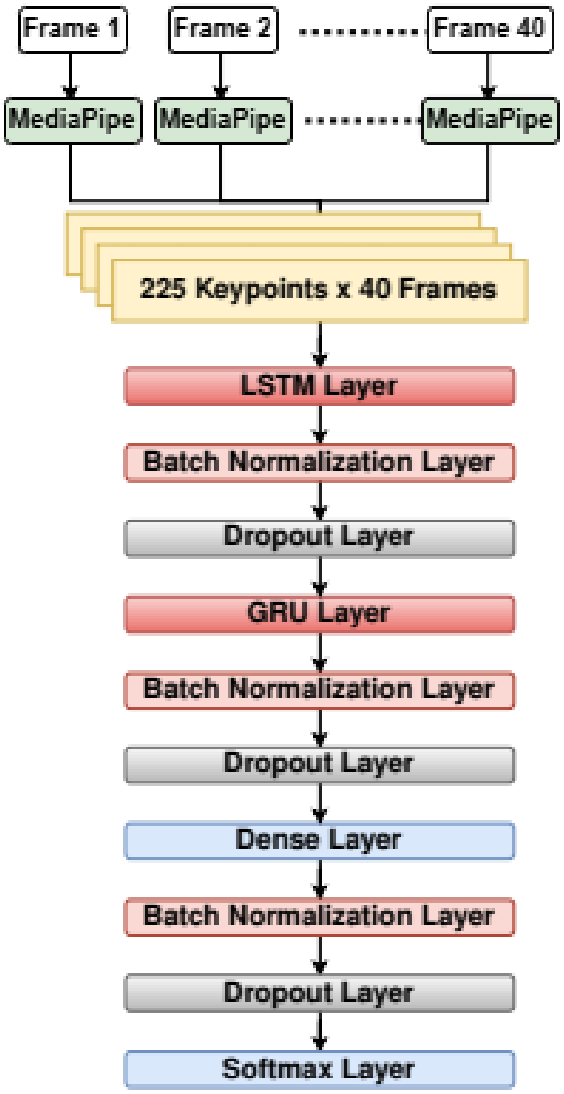

Fig. 4. Hybrid LSTM-GRU model architecture proposed by Navendu and Sahula [2024]

Navendu and Sahula [2024] proposed a hybrid LSTM-GRU network, leveraging the keypoints of video frames extracted through MediaPipe Hands and Pose for feature extraction. In line with that, 225 keypoints, describing hand and body landmarks, were extracted out of every frame to be processed. The hybrid model combines LSTM and Gated Recurrent Unit (GRU) layers as seen in Figure 4, to model the temporal dependencies that could exist within the

gestures. It was tested on the publicly available ISL dataset INCLUDE, and it reported 89.5% accuracy.

2.4 Arabic Sign Language (ArSL)

AL-Rousan et al. [2009] introduced a basic recognition system using Hidden Markov Model (HMM) and a self-collected dataset of 30 isolated words with 7,860 gestures recorded at a framerate of 25fps from 18 signers. Discrete Cosine Transform (DCT) was used for feature extraction and then Zonal coding was applied, followed by HMM classification, achieving 90.6% in online signer-independent mode and 97.4% accuracy in offline signer-dependent mode.

DeepArSLR is a framework introduced by Aly and Aly [2020]. As seen in Figure 5, DeepLabv3+ was used for precise hand segmentation which accurately extracts hand regions from video frames and Convolutional Self-Organising Map (CSOM) was implemented to capture detailed hand shape features. Temporal features were modeled using a three-layer Bi-LSTM network to learn the sequential nature of the gestures. DeepArSLR was tested and worked upon the ArSL database collected in [Shanableh et al. 2007a]. The framework achieved a respectable accuracy of 89.5%.

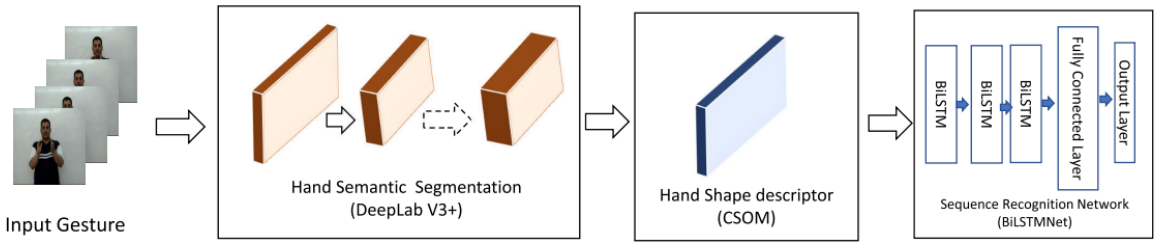

Fig. 5. Proposed DeepArSLR framework by Aly and Aly [2020]

Similar to AL-Rousan et al. [2009], Sidig et al. [2021]. used HMM for Classification. Histogram of Oriented Gradients (HOG) and skeleton joint coordinates that were obtained from the Kinect sensor were used to extract features. Sidig et al. [2021] presented a new large-scale KArSL dataset consisting of 502 signs repeated 50 times by 3 professional signers resulting in 75300 samples. The system achieved an overall accuracy of 89%. However, the accuracy dropped significantly for signer-independent scenarios.

Alyami et al. [2024] used a subset of 100 signs from Sidig et al. [2021]'s KArSL dataset to propose a transformer-based model using a combination of hand and face key points extracted with the MediaPipe pose estimator. Three models were explored LSTM, Temporal Convolution Networks (TCN), and Transformer where the latter showed the comparative best performance

due to its self-attention mechanism that effectively captured the complex dependencies between gestures. The framework achieved a remarkable accuracy of 99.74% in signer-dependent mode and 68.2% in signer-independent mode outperforming other state-of-arts models on KArSL-100 dataset at the time.

2.5 American Sign Language (ASL)

Li et al. [2020] proposed a Pose-based Temporal Graph Convolutional Network (TGCN) by modeling both spatial as well as temporal dependencies within keypoint sequences. The model was trained on WLASL dataset introduced by authors themselves and comprising 21,083 videos of 119 individuals performing 2,000 signs. OpenPose was employed to capture 55 body and hand keypoints. The TGCN achieved 23.65% top-1 accuracy on the WLASL2000 subset, while a fine-tuned I3D model performed slightly better with 32.48% top-1 accuracy. Li et al. [2020] note that even though I3D is larger than the proposed Temporal Graph Convolutional Network (TGCN), pose-Temporal Graph Convolutional Network (TGCN) achieves comparable top-5 and top-10 accuracy to Two-Stream Inflated 3D ConvNet (I3D) on WLASL2000.

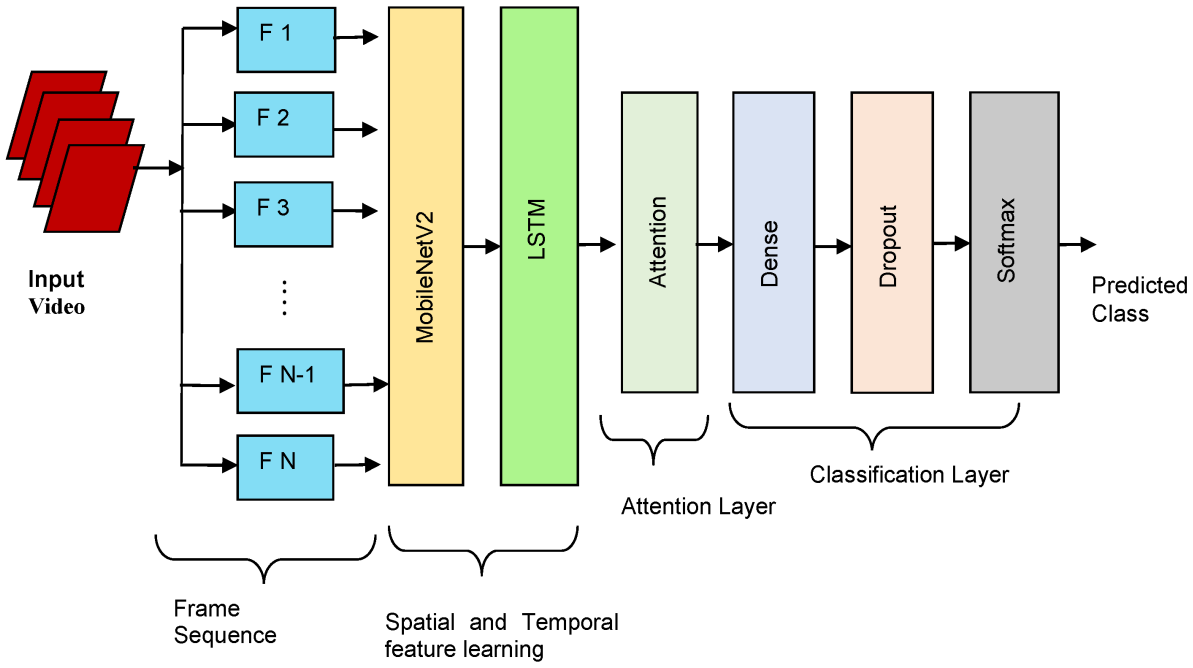

Fig. 6. Block diagram of the proposed approach by [Kumari and Anand 2024]

Building on the same dataset, Kumari and Anand [2024] proposed a hybrid CNN-LSTM framework integrated with an attention mechanism, as shown in Figure 6. This model was trained on a subset of the WLASL dataset [Li et al. 2020]. Kumari and Anand [2024] employed

a pre-trained MobileNetV2 model to extract spatial features from video data. Afterwards, these features were subsequently processed by LSTM layers with an enhanced attention mechanism. This allows the model to pay attention to relevant information about hand gestures throughout the time and yielded an accuracy of 84.65%.

2.6 Korean Sign Language (KSL)

Shin et al. [2023b] proposed a multi-branch architecture combining CNN and Transformer modules for KSL recognition. The reliability of this approach was demonstrated through experiments on the KSL-77 dataset [Yang et al. 2020] and their proposed dataset KSL-20. In this approach, grain architecture has been used to extract fine features from the beginning, followed by the parallel feature extraction through CNN for local features and transformer for capturing long-range dependencies and is finally classified through concatenation by means of a module that includes global average pooling and a fully connected layer. The model achieved a respectable accuracy of 89.00% for the KSL-77 dataset and a much higher impressive one, 98.30%, in the proposed dataset.

Building upon their previous work, Shin et al. [2023a] extended their research on KSL recognition using both KSL-77 and their self-collected dataset, KSL-20. Compared with their earlier work, While the used approach was based on a multi-branch CNN-Transformer in the previous paper, this advanced paper represents a two-stream deep learning net combined with Graph Convolutional Network (GCN) and attention-based neural networks. It proposes taking 47 pose landmarks from videos using MediaPipe and feeding them into the proposed model. In summary, one stream captures the spatial features of the appears, while another stream focuses on joint motion. Proper refinement steps, including channel attention and a general CNN, will be done at both. The performance of the model has reached a remarkable accuracy of 99.87% for the KSL-77 dataset, as well as 100.00% on the KSL-20 dataset.

2.7 Critical Analysis

The most notable gap identified from our study is the lack of standardized and robust PSL datasets such as those available for ASL, by Li et al. [2020], ISL [Sridhar et al. 2020], which limits the generalizability and scalability of PSL-focused models. Research on PSL, including the work by Sameena Javaid [2023], often uses small, non-standardized datasets such as the PkSLMNM dataset [Javaid 2022], which limits the real-world performance of models.

The answer to this problem lies in large and standardized PSL datasets and efficient data augmentation techniques for improving model robustness with lesser overfitting under various user conditions. Additionally, ISL studies [Adithya and Rajesh 2020; Aparna and Geetha 2020; Das et al. 2023; Mittal et al. 2019; Navendu and Sahula 2024] and ASL research [Kumari and Anand 2024; Li et al. 2020] employ deep learning temporal-spatial models such as CNNs and LSTM. While these complex architectures are rather standard for dynamic gesture recognition, PSL studies tend to be carried out using simpler architectures that inadequately capture the subtle spatial and temporal aspects of PSL. For example, Although Sameena Javaid [2023] applies I3D-ShuffleNet for feature extraction, the approach lacks adaptive temporal modeling provided by Transformers or even Bi-LSTM networks observed in other research [Alyami et al. 2024; Shin et al. 2023a,b]. This difference in feature extraction and classification approaches suggest a limitation of the current PSL frameworks in relation to this issue. These include its ability to accurately capture gesture dynamics, emphasizing a potential for integrating hybrid architectures, a limitation addressed by Kumari and Anand [2024], where they combine spatiotemporal models to enhance the recognition accuracy.

Another critical gap includes the fact that dynamic texture descriptors have not been explored in PSL recognition. Dynamic texture descriptors are used for capturing motion patterns of video sequences but have not been widely applied in existing PSL studies. Also, the possible fusion of handcrafted descriptors with deep learning-based models has equally been overlooked. On one hand, handcrafted descriptors like Motion Binary Pattern (MBP) can give rich and complementary temporal information; on the other hand, deep learning models are very good at extracting high-level spatial features. In this framework, therefore, such an MBP is chosen to carry out the temporal feature extraction while ResNet-101 does the spatial features to allow us to perform a hybrid of both hand-crafted and deep learning-based approaches for enhanced PSL recognition. These features will then be fed to transformer, Bi-LSTM and GRU models. A majority vote will be employed to give prediction.

2.8 Comparison of Related Works

A comparison of all relevant studies was made. Table 1 shows an illustrative comparison for all the mentioned frameworks presented by authors, the dataset used, and type of data. It also shows which algorithm was used for feature extraction, as well as for classification. Finally, the accuracy is presented as a way to show how the frameworks performed after a test run using the dataset.

References	Domain	Dataset	Feature Extraction	Classification	Accuracy (%)
Sameena Javaid [2023]	PSL	PkSLMNM [Javaid 2022]	Spatiotemporal features with I3D-ShuffleNet	Action Transformer with RPN	86.12
Mirza et al. [2022]	PSL	Author collected data	SURF algorithm + Bag-of-Words Model	SVM	96.53
Hamza and Wali [2023]	PSL	Subset of PSL Dictionary [psl.org.pk 2020]	C3D, I3D, TSM	C3D	93.33
Mittal et al. [2019]	ISL	Author collected data	2D-CNN	Modified 4 Gated LSTM	89.50
Aparna and Geetha [2020]	ISL	Author collected data	Inception V3 (pretrained CNN)	Stacked LSTM	94.00
Adithya and Rajesh [2020]	ISL	Emergency Words [V and R 2021]	3D wavelet transform descriptors	GoogleNet + LSTM	96.25
Das et al. [2023]	ISL	Emergency Words [V and R 2021]	SIFT + BOVW + VGG-19	Bi-LSTM	94.42
Navendu and Sahula [2024]	ISL	INCLUDE [Sridhar et al. 2020]	Keypoints with MediaPipe Hands and Pose	Hybrid LSTM-GRU	89.50
AL-Rousan et al. [2009]	ArSL	Author collected data	DCT	HMM	97.40
Aly and Aly [2020]	ArSL	ArSL Database [Shanableh et al. 2007b]	DeepLab v3+ + CSOM	Bi-LSTM	89.50
Sidig et al. [2021]	ArSL	KArSL-100 [Sidig et al. 2021]	HOG + Skeleton joint coordinates	HMM	89.00
Alyami et al. [2024]	ArSL	KArSL-100 [Sidig et al. 2021]	2D pose landmarks of hands and face	Transformer	99.74
Li et al. [2020]	ASL	WLASL [Li et al. 2020]	OpenPose (55 keypoints)	Pose-based TGCN	23.65 (top-1)
Kumari and Anand [2024]	ASL	Subset of WLASL [Li et al. 2020]	MobileNetV2	LSTM with Attention	84.65
Shin et al. [2023b]	KSL	KSL-20 [Shin et al. 2023b]	Grain architecture with CNN and Transformer	Multi-branch CNN-Transformer	98.30
Shin et al. [2023a]	KSL	KSL-20 [Shin et al. 2023b]	MediaPipe Pose Landmarks + GCN	Two-stream GCN with Attention	100.00

Table 1. Comparison Table of Related Works

3 METHODOLOGY

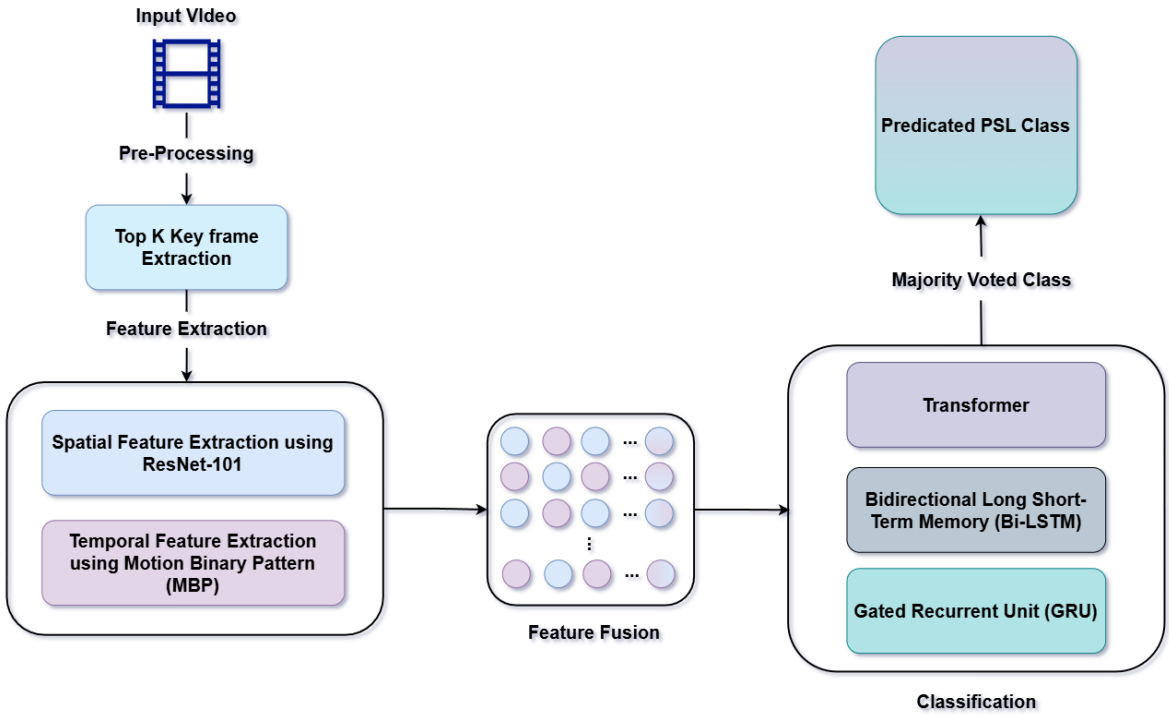

Fig. 7. Basic Architecture of the proposed method

As seen in Figure 7, the proposed framework is a two-stream approach using video data to extract PSL signs from the PkSLMNM dataset, which are pre-defined into categorized videos on specified signs. The frames are selected through Top-K Key Frame Extraction [Joolee et al. 2018] algorithm through the computation of histogram differences to select the ones that best represent the most relevant dynamic information, reducing the computational cost with a minimum loss of key information. After that, these selected key frames will be resized to a standard size of 224x224 pixels and then converted into grayscale.

To grasp the detailed spatial patterns for videos, ResNet-101 [He et al. 2016] is leveraged, whose major deep residual learning framework can ensure feature extraction with minimized vanishing gradient problems for very deep networks. To catch motion information, handcrafted descriptor MBP [Baumann et al. 2014] is used, by analyzing pixel intensity variations between consecutive frames and encoding these as binary patterns, temporal features are obtained.

The features extracted are combined and then fed into a series of models for classification. Transformers [Vaswani et al. 2017] are synonymous for their attention mechanisms and are used to extract long-range dependencies and relationships in sequential data, further enhancing the contextual understanding of the PSL signs by the model. A Bi-LSTM [Graves and Schmidhuber 2005] is a bidirectional extension of the LSTM [Hochreiter and Schmidhuber 1997] which captures comprehensive temporal features, representing temporal dependencies in both forward and backward temporal directions. Gated Recurrent Unit (GRU) [Cho et al. 2014] is efficient yet maintains performance for handling long-term dependencies in sequential data.

The final classification is obtained through the majority voting mechanism, allowing the best overall predictions by these models to enhance both accuracy and robustness. This integrated approach creates a robust system capable of achieving a rich representation of PSL gestures, improving overall recognition accuracy and adaptability to real-world conditions.

4 REQUIREMENT ANALYSIS

In this section we will see the main requirements for our model based on MoSCoW framework, the following requirements are classified as Must have, Should have, Could have, or Won't have. It is divided into three sections where functional, non-functional, and hardware requirements are shown in Table 2, Table 3, and Table 4 respectively.

4.1 Functional Requirements

ID	Requirement Description	Priority
FR1	Split the PkSLMNM dataset into appropriately sized training and testing subsets	Must
FR2	Extract key frames from input videos using Top K Key frame extraction before being fed to the model	Must
FR3	Extract spatial features using ResNet-101	Must
FR4	Extract temporal motion information using MBP technique	Must
FR5	Apply feature fusion to concatenate the extracted features.	Must
FR6	Implement three sequential models (Transformers, Bi-LSTM and GRU) individually	Must
FR7	The final prediction must be generated based on majority voting	Must
FR8	Use the test subset to evaluate the performance of the models	Must
FR9	Perform hyperparameter optimization on the models	Must
FR10	Save the preprocessed data within the same directory for ease of use	Should
FR11	Save the extracted spatial and temporal features to a separate location locally for possible reusability and re-training	Should
FR12	Implement progress bars and debugging statements for readability	Could
FR13	Produce intermediate outputs for each model, such as attention weights or hidden states, to aid in interpretability	Could

Table 2. Functional Requirements Table

4.2 Non-Functional Requirements

ID	Requirement Description	Priority
NFR1	The model must achieve a respectable 85% or more accuracy for PSL gestures across the framework	Must
NFR2	Data must be stored locally and processed to ensure no potential data leak	Must
NFR3	The framework should perform consistently with minimal variation across the different signs	Should
NFR4	Access to the dataset, model configurations, and outputs should be restricted to the author and supervisor only	Should
NFR5	Model can be used to support additional PSL signs without significant restructuring	Could

Table 3. Non-Functional Requirements Table

4.3 Hardware & Software Requirements

Here are the hardware & Software requirements needed for this project. Since we are working with video data, the computational performance will be demanding. Hence, to ensure efficient processing and model accuracy, the following hardware specifications are recommended.

ID	Component	Description
HR1	Processor (CPU)	AMD Ryzen 7, optimized for multi-threaded performance
HR2	Graphics Card (GPU)	CUDA-cores supported NVIDIA RTX 3050 GPU, 4 GB VRAM
HR3	Memory (RAM)	32 GB DDR4
HR4	Storage	2 TB SSD
HR5	Software Compatibility	Supports TensorFlow and Keras back-end to leverage GPU acceleration for efficient model training

Table 4. Hardware & Software Requirements Table

5 EVALUATION

5.1 Dataset

For this research, we will train and evaluate the proposed method with a single dataset. By using the Pakistan Sign Language Manual and Non-Manual (PkSLMNM) dataset [Sameena Javaid 2023], which is specifically designed for Pakistan Sign Language (PSL) identification and includes both manual as well as non-manual gestures. The dataset consists of 665 videos of 180 people, which include 70 females and 110 males, which vary in age from 20 to 50 years old as below in Figure 8.

Fig. 8. Samples of the PkSLMNM dataset

The PkSLMNM dataset includes a range of PSL expressions represented by facial and hand gestures, covering seven emotional categories as seen in the Table 5. Each video is recorded in HD at 1920x1080 resolution, lasting approximately 2 seconds per sample. The recordings were made at a frame rate of 25 frames per second (fps).

Sign	Number of Samples
Bad	97
Best	98
Glad	98
Sad	95
Scared	94
Stiff	85
Surprise	98

Table 5. Samples per sign in the PkSLMNM dataset

5.2 Evaluation Metrics

In this section we will be exploring the following performance metrics that will be used to evaluate our model:

(1) Accuracy;
(2) Precision;
(3) Recall;

5.2.1 Accuracy. :

Accuracy is defined as the proportion of correctly predicted signs to the total number of predictions made. Usually, it would represent the performance of the model in the general success rate of the recognition of signs. However, this could not be useful as a metric when dealing with an imbalanced dataset. Mathematically, accuracy is represented as:

$$Accuracy = \frac{TruePositives(TP) + TrueNegatives(TN)}{TruePositives(TP) + TrueNegatives(TN) + FalsePositives(FP) + FalseNegatives(FN)}$$

where

- **True Positives (TP):** The number of PSL signs model predicted accurately.
- **True Negatives (TN):** The number of non-PSL signs model predicted accurately.
- **False Positives (FP):** The number of non-PSL signs model predicted inaccurately as PSL signs.
- **False Negatives (FN):** The number of PSL signs model predicted inaccurately as non-PSL signs.

5.2.2 Precision. :

Precision estimates the capability of a model to avoid false positives. It calculates the ratio of the correctly predicted positive signs to the total number of predicted positive signs. With a higher precision, it means that the model will be reliable in classifying PSL signs without doing so in a faulty manner where the non-signs are interpreted as signs. High precision indicates that the model's positive predictions are right and it doesn't mistakenly classify any non-signs as signs, which is crucial for the reduction of miscommunication in conditions of PSL recognition. Precision is calculated as follows:

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)}$$

5.2.3 *Recall.* :

Recall, or sensitivity, is the proportion of relevant PSL signs in the test subset that the model can detect gestures. It gives the ratio of correctly predicted positive signs concerning the total actual positive PSL signs are reflective of the model performance in capturing the PSL signs when present. Recall is computed as:

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)}$$

A high recall score means that most PSL signs have been successfully recognized by the model, with only a few of them being misclassified. This contributes to ensuring that few if any relevant gesture sequences will be overlooked. Balancing recall with precision is important because it makes sure the model captures the signs accurately and identifies as many relevant signs as possible.

6 CONCLUSION

In conclusion, this dissertation aims to fill in the gaps in Pakistan Sign Language Recognition (PSLR), proposing a robust framework that shall be integrated with the advanced models and techniques. As a key method of communication for Pakistan's deaf and hard-of-hearing community, PSL has remained under-investigated due to the unavailability of standardized datasets and comprehensive methodologies. An extensive review of related work (refer to Section 3) under PSL and related domains such as American, Indian, Arabic, and Korean Sign Language was done to find gaps in the literature.

Based on these findings, we aim to propose a two-stream approach (refer to Section 3) using ResNet-101 for appearance spatial feature extraction, and MBP for temporal analysis. Classification would be done through a majority voting mechanism among advanced models such as Transformers, Bi-LSTM, and GRU. This methodology solves the challenge of capturing the spatial as well as temporal complexities of PSL signs while maintaining scalability and reliability across various kinds of user conditions.

While this framework looks promising, its implementation remains a focus for our future work. Additionally, the results shall be compared to other approaches in the field of Sign Language Recognition.

REFERENCES

V. Adithya and R. Rajesh. 2020. Hand gestures for emergency situations: A video dataset based on words from Indian sign language. *Data in Brief* 31 (Aug. 2020), 106016. https://doi.org/10.1016/j.dib.2020.106016

M. AL-Rousan, K. Assaleh, and A. Tala'a. 2009. Video-based signer-independent Arabic sign language recognition using hidden Markov models. *Applied Soft Computing* 9, 3 (June 2009), 990–999. https://doi.org/10.1016/j.asoc.2009.01.002

Saleh Aly and Walaa Aly. 2020. DeepArSLR: A Novel Signer-Independent Deep Learning Framework for Isolated Arabic Sign Language Gestures Recognition. *IEEE Access* 8 (2020), 83199–83212. https://doi.org/10.1109/ACCESS.2020.2990699

Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. Isolated Arabic Sign Language Recognition Using a Transformer-based Model and Landmark Keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, 1 (Jan. 2024), 1–19. https://doi.org/10.1145/3584984

C. Aparna and M. Geetha. 2020. CNN and Stacked LSTM Model for Indian Sign Language Recognition. In *Machine Learning and Metaheuristics Algorithms, and Applications*, Sabu M. Thampi, Ljiljana Trajkovic, Kuan-Ching Li, Swagatam Das, Michal Wozniak, and Stefano Berretti (Eds.). Springer, Singapore, 126–134. https://doi.org/10.1007/978-981-15-4301-2_10

Florian Baumann, Jie Lao, Arne Ehlers, and Bodo Rosenhahn. 2014. Motion Binary Patterns for Action Recognition. In *International conference on pattern recognition applications and methods*, Vol. 2. SCITEPRESS, 385–392. https://doi.org/10.5220/0004816903850392

Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, action recognition? a new model and the Kinetics dataset. https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html

Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. 2019. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics* 10, 1 (Jan. 2019), 131–153. https://doi.org/10.1007/s13042-017-0705-5

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078* (Sept. 2014). https://doi.org/10.48550/arXiv.1406.1078

Soumen Das, Saroj Kr Biswas, and Biswajit Purkayastha. 2023. Automated Indian sign language recognition system by fusing deep and handcrafted feature. *Multimedia Tools and Applications* 82, 11 (May 2023), 16905–16927. https://doi.org/10.1007/s11042-022-14084-4

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5–6 (July 2005), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Hafiz Muhammad Hamza and Aamir Wali. 2023. Pakistan sign language recognition: leveraging deep learning models with limited dataset. *Machine Vision and Applications* 34, 5 (July 2023), 71. https://doi.org/10.1007/s00138-023-01429-8

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation MIT-Press* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Sameena Javaid. 2022. PkSLMNM: Pakistan sign language manual and non-manual gestures dataset. https://data.mendeley.com/datasets/m3m9924p3v/2

Joolekha Joolee, Md Uddin, Jawad Khan, Taeyeon Kim, and Young-Koo Lee. 2018. A Novel Lightweight Approach for Video Retrieval on Mobile Augmented Reality Environment. *Applied Sciences* 8, 10 (Oct. 2018), 1860.

https://doi.org/10.3390/app8101860

Binwant Kaur, Aastha Chaudhary, Shahina Bano, Yashmita, S.R.N. Reddy, and Rishika Anand. 2023. Fostering inclusivity through effective communication: Real-time sign language to speech conversion system for the deaf and hard-of-hearing community. *Multimedia Tools and Applications* 83, 15 (Oct. 2023), 45859–45880. https://doi.org/10.1007/s11042-023-17372-9

Diksha Kumari and Radhey Shyam Anand. 2024. Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on Attention Mechanism. *Electronics* 13, 77 (March 2024), 1229. https://doi.org/10.3390/electronics13071229

Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Snowmass Village, CO, USA, 1448–1458. https://doi.org/10.1109/WACV45572.2020.9093512

Muhammad Shaheer Mirza, Sheikh Muhammad Munaf, Fahad Azim, Shahid Ali, and Saad Jawaid Khan. 2022. Vision-based Pakistani sign language recognition using bag-of-words and support vector machines. *Scientific Reports* 12, 1 (Dec. 2022), 21325. https://doi.org/10.1038/s41598-022-15864-6

Anshul Mittal, Pradeep Kumar, Partha Pratim Roy, Raman Balasubramanian, and Bidyut B. Chaudhuri. 2019. A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion. *IEEE Sensors Journal* 19, 16 (Aug. 2019), 7056–7063. https://doi.org/10.1109/JSEN.2019.2909837

Kumar Navendu and Vineet Sahula. 2024. Word Level Sign Language Recognition using MediaPipe and LSTM-GRU Network. *Authorea Preprints* (July 2024). https://doi.org/10.36227/techrxiv.172054945.57389794/v1

World Health Organization et al. 2021. *World report on hearing* (1st ed ed.). World Health Organization, Geneva.

PADEAF. [n. d.]. Deaf Statistic | PADEAF. https://www.padeaf.org/quick-links/deaf-statistics

psl.org.pk. 2020. https://www.psl.org.pk

Safdar Rizvi Sameena Javaid. 2023. A Novel Action Transformer Network for Hybrid Multimodal Sign Language Recognition. *Computers, Materials & Continua* 74, 1 (2023), 523–537. https://doi.org/10.32604/cmc.2023.031924

Syed Muhammad Saqlain Shah, Javed I. Khan, Syed Husnain Abbas, and Anwar Ghani. 2023. Symmetric mean binary pattern-based Pakistan sign language recognition using multiclass support vector machines. *Neural Computing and Applications* 35, 1 (Jan. 2023), 949–972. https://doi.org/10.1007/s00521-022-07804-2

Tamer Shanableh, Khaled Assaleh, and Mohammad Al-Rousan. 2007a. Spatio-temporal feature-extraction techniques for isolated gesture recognition in Arabic sign language. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, 3 (2007), 641–650.

Tamer Shanableh, Khaled Assaleh, and M. Al-Rousan. 2007b. Spatio-Temporal Feature-Extraction Techniques for Isolated Gesture Recognition in Arabic Sign Language. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, 3 (June 2007), 641–650. https://doi.org/10.1109/TSMCB.2006.889630

Jungpil Shin, Abu Saleh Musa Miah, Kota Suzuki, Koki Hirooka, and Md. Al Mehedi Hasan. 2023a. Dynamic Korean Sign Language Recognition Using Pose Estimation Based and Attention-Based Neural Network. *IEEE Access* 11 (2023), 143501–143513. https://doi.org/10.1109/ACCESS.2023.3343404

Jungpil Shin, Abu Saleh Musa Miah, Md. Al Mehedi Hasan, Koki Hirooka, Kota Suzuki, Hyoun-Sup Lee, and Si-Woong Jang. 2023b. Korean Sign Language Recognition Using Transformer-Based Deep Neural Network. *Applied Sciences* 13, 55 (Feb. 2023), 3029. https://doi.org/10.3390/app13053029

Seongjoo Shin, Youngmi Baek, Jinhee Lee, Yongsoon Eun, and Sang Hyuk Son. 2017. Korean sign language recognition using EMG and IMU sensors based on group-dependent NN models. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–7. https://doi.org/10.1109/SSCI.2017.8280908

Ala Addin I. Sidig, Hamzah Luqman, Sabri Mahmoud, and Mohamed Mohandes. 2021. KArSL: Arabic Sign Language Database. *ACM Transactions on Asian and Low-Resource Language Information Processing* 20, 1 (Jan. 2021), 14:1–14:19. https://doi.org/10.1145/3423420

Venus Singla, Seema Bawa, and Jasmeet Singh. 2024. Enhancing Indian sign language recognition through data augmentation and visual transformer. *Neural Computing and Applications* 36, 24 (Aug. 2024), 15103–15116. https://doi.org/10.1007/s00521-024-09845-1

Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 1366–1375. https://doi.org/10.1145/3394171.3413528

Adithya V and Rajesh R. 2021. A Video Dataset of the Hand Gestures of Indian Sign Language Words used in Emergency Situations. *Data in Brief* 1 (Aug. 2021). https://doi.org/10.17632/2vfdm42337.1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need, In Advances in Neural Information Processing Systems. *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Seunghan Yang, Seungjun Jung, Heekwang Kang, and Changick Kim. 2020. The Korean Sign Language Dataset for Action Recognition. In *MultiMedia Modeling*, Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve (Eds.). Springer International Publishing, Cham, 532–542. https://doi.org/10.1007/978-3-030-37731-1_43

A PROJECT MANAGEMENT

This section highlights our project plan toward this study which includes the setting of milestones, and monitoring progress towards the attainment of objectives. We have organized the plan around key deliverables associated with timelines for timely completion in the form of gantt charts. In addition, we have identified some of the risks that could be associated with this project and proactively suggest mitigation strategies as a table. The comprehensive approach shall guide the project in its various phases while focusing on quality and objective accomplishment.

A.1 Project Scope

The scope of this project goes toward the development and proposal of a reliable framework for Pakistan Sign Language (PSL) recognition by addressing significant gaps in PSL research, for example the limited use of hybrid methodologies and algorithms that have not been used in the realm of sign language recognition, particularly PSL. This architecture design emphasizes the recognition of PSL signs with high accuracy from the video data, taking a pre-processing step through Top-K Key Frame Extraction, followed by spatial feature extraction with ResNet-101, while temporal analysis is achieved via MBP. Sequential classification is realized through transformer, Bi-LSTM, and GRU. In the end, result is predicted with ensembling through majority voting to keep this model design robust, interpretable, and scalable. This research will contribute to the advancement of recognizing PSL and help in promoting communication within the deaf and hard-of-hearing community in Pakistan.

A.2 Project Deliverables

This project is divided into a set of deliverables, which are to be delivered at their respective deadlines. All said tasks will be completed in a timely and organized manner for the completion of the project. The deliverables are listed below:

- Research Report (Semester 1): The first deliverable is the Research Report containing the fundamental layout at the start of the project. This report shall contain the aims and objectives of the project, literature review of related works, detailed requirement analysis, overview of the methodology, lays down an approach toward evaluation, and gives a preliminary timeline for the project in order to set directions for subsequent work. This report also addresses professional, legal, ethical and social issues.

- Dissertation Report (Semester 2): The second deliverable is a comprehensive dissertation that documents the entire lifecycle of the project. It shall contain an in-depth description of the methodology used, the evaluation results, and a discussion of the findings in light of related work while also highlighting the project's contributions and areas for future research.

- Viva/Poster Presentation: The last deliverables of the project are the poster and mini-viva. The poster is a simple overview of the project, a description of the main objectives, methodologies, results, and conclusions. The mini-viva extends this by allowing an in-depth discussion and explanation of the work to be presented, questions to be answered, and the significance of the work to be outlined. Each contributes to another in presenting the work both effectively and easily understood.

A.3 Project Plan

The project plan accounts for the main tasks, deadlines, and also milestones. This is shown via gantt charts in Figure 9 and Figure 10 which depicts them in phases, starting with the initialization of the project right through to the end submission for both deliverables. The first chart covers the first deliverable, and all the tasks that were covered and will be covered during that period will be listed accordingly for easy view. The second chart will be an approximation of how things are meant to come along according to set time period for each task for the next semester.

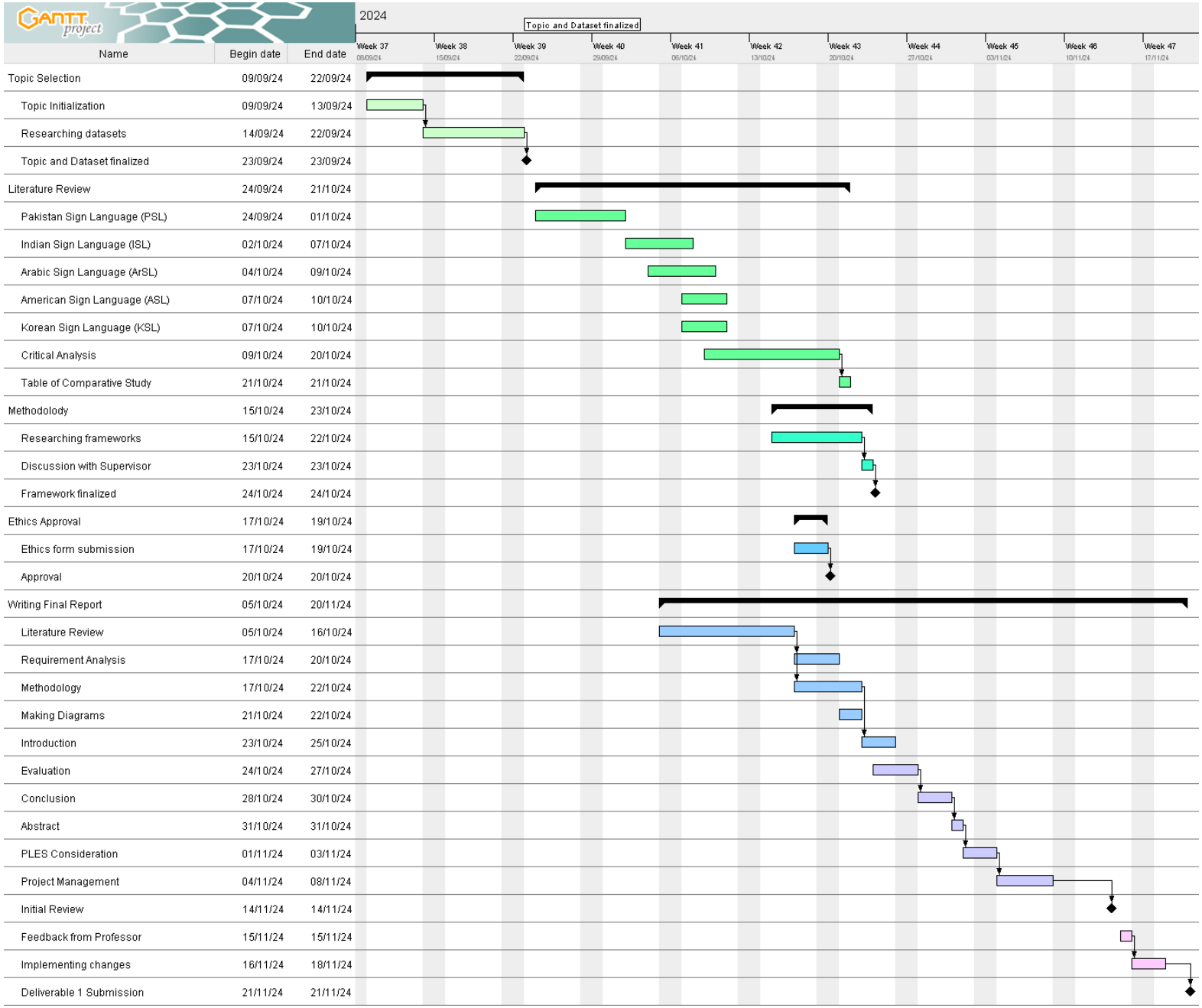

Fig. 9. Gantt Chart for Semester 1

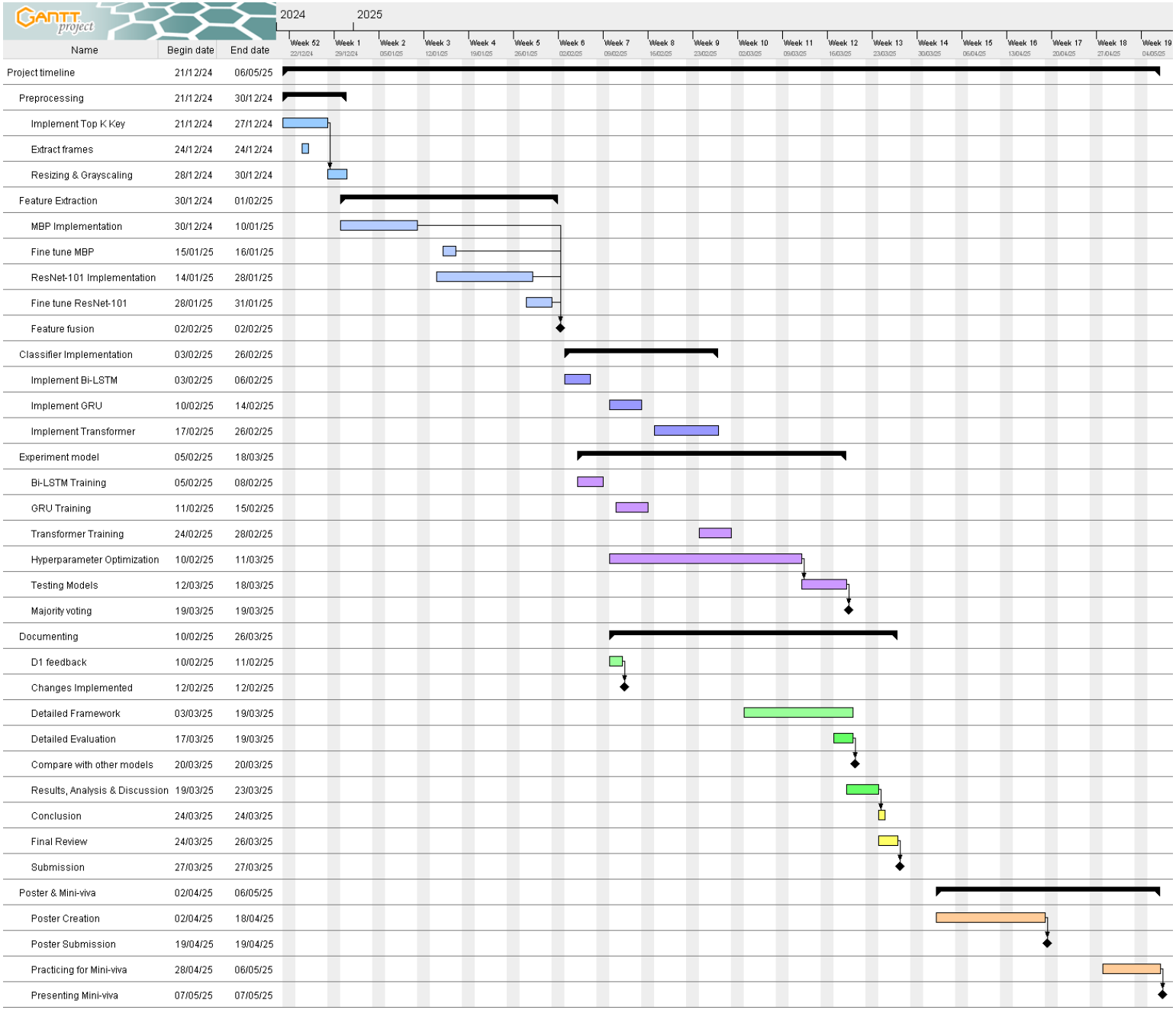

Fig. 10. Gantt Chart for Semester 2

A.4 Risk Analysis

Risk analysis is a vital component in the management of a project, as it locates those potential risks that might impede the successful completion of the project. For each risk identified, two main dimensions have been considered: the likelihood, relating to the opportunity that the risk will occur, and the impact, which describes the magnitude of the damage if the risk actually happens. These can be combined to provide a means of categorizing risks so that mitigation efforts can be targeted effectively. For instance, highly likely risks with a high impact-threats such as overfitting due to limited data require an immediate and intense mitigation strategy. A low-likelihood risk that does have a high impact, such as the failure of a system, requires the implementation of precautionary measures beforehand, for instance through periodic backups, to limit damage. Table 6 below summarizes key identified risks for this project, their likelihood, impact, and corresponding mitigation strategies.

Risk	**Likelihood**	**Impact**	**Mitigation Strategy**
Main system(s) stops working	Unlikely	Very High	Regularly back up work to multiple locations (external devices or cloud storage).
Data loss or corruption	Unlikely	Very High	Keep track of progress by using version control.
Overfitting due to limited data	Likely	High	Implement data augmentation techniques and cross-validation.
Incompatibility between software tools	Possible	Medium	Test the software during initial setup extensively and use proper versions.
Ethics form gets disapproved	Unlikely	High	Communicate with supervisor to ensure validity of the ethics form.
Low performance of the proposed framework	Likely	High	Fine-tune the hyperparameters or increase the training set size.

Table 6. Risk Analysis Table

B PROFESSIONAL, LEGAL, ETHICAL AND SOCIAL ISSUES

B.1 Professional Issues

The proposed framework will be implemented using Python and by using only licensed open-source software and libraries some of which include TensorFlow, Pandas and Keras. Any models or libraries used within the research will be correctly referenced with regard to their terms and conditions. This project follows the code of conduct as specified by institutions such as the British Computing Society.

B.2 Legal Issues

This project does not store any personal information or any user data. It strictly follows data protection laws and regulations of the United Arab Emirates, Pakistan's Personal Data Protection Act (PDPA) and relevant General Data Protection Regulation (GDPR) guidelines. The data is stored locally and under the rules and guidelines of the university and will not be misused in any way. The relevant files will not be distributed to any system other than the author's system(s). The proposed framework does not and will not violate any laws.

B.3 Ethical Issues

In this research based study we will only use the publicly available PkSLMNM dataset introduced by Sameena Javaid [2023] and will not have any human subjects. We will not be involving any external human user-based interaction or feedback during this project. The performance of the framework will be only validated on empirical data alone. Each sign category will be treated equally during training to avoid bias. The study ensure that no ethical policies and principles are violated in any way shape or form.

B.4 Social Issues

The purpose of this study is to develop a sign language recognition system specifically for PSL. Since this project is strictly technical, it does not introduce any sensitive or controversial social issues. The project aims to enhance PSL recognition to support the deaf and hard-of-hearing community without any potential for misuse or ethical conflicts. It is important to note that this research does not address, nor does it introduce, any broader social or moral issues.