Real-Time Bangladeshi Sign Language Recognition and Natural Language Interpretation Using Hybrid Deep Learning and LLM

by

Fabliha Akther Fairuz
22101200
Santonu Roy
21201475
Sajid Mahmud
20101076
Sumiya Tasnim
21201518
Inkiad Bin Ershad Rafey
21201516

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
June 2026.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing our degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Fabliha Akther Fairuz
22101200

Santonu Roy
21201475

Sajid Mahmud
20101076

Sumiya Tasnim
21201518

Inkiad Bin Ershad Rafey
21201516

Approval

The thesis/project titled "Real-Time Bangladeshi Sign Language Recognition and Natural Language Interpretation Using Hybrid Deep Learning and LLM" submitted by

1. Fabliha Akther Fairuz (22101200)
2. Santonu Roy (21201475)
3. Sajid Mahmud (20101076)
4. Sumiya Tasnim (21201518)
5. Inkiad Bin Ershad Rafey (21201516)

of Summer 2026 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science & Engineering on June 2026.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science & Engineering
BRAC University

Thesis Coordinator:
(Member)

Md Golam Rabiul Alam, PhD
Professor
Department of Computer Science & Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson, Associate Professor
Department of Computer Science & Engineering
Brac University

Dedication

To the Deaf and hard-of-hearing community of Bangladesh,
whose voices inspired every line of this work.

Ethics Statement

All participants in the data collection process provided informed consent consent for the use of their recordings in this thesis and for illustrative academic purposes. The vocabulary and expression categories were validated by a certified representative of the Deaf community, who also consented to having his name and designation cited. No identifiable video of any participant is stored or transmitted by the system; only skeletal landmark coordinates are processed during training and inference. This project did not require a formal institutional ethics board approval, but all data collection and usage followed the ethical principles of informed consent, participant privacy, and responsible AI development as described in Chapter 3.

Acknowledgements

The authors thank Mohammad Arish Abedin, Life Member of the Bangladesh National Federation of the Deaf and Vice President of the Bangladesh Student Welfare Federation of the Deaf, for reviewing the dataset vocabulary and expression categories from the perspective of a native BdSL user. His confirmation that the selected signs are widely used within the Deaf community and that the five expression categories are appropriate as non-manual markers in BdSL strengthened the linguistic grounding of this work. He also provided consent for his name and designation to be cited.

Sincere gratitude is extended to supervisor Dr. Md. Ashraful Alam for his guidance throughout the research process.

Abstract

Deaf and hard-of-hearing communities in Bangladesh face a persistent communication gap: existing sign language recognition systems process hand gestures but ignore the facial expressions and head movements that carry grammatical meaning in Bangladeshi Sign Language (BdSL). This study introduces BdSL-NMM, the first BdSL dataset annotated with non-manual marker labels alongside word-level glosses, covering 62 sign classes and 5 expression categories across 4,170 recordings from seven native signers. To evaluate on this dataset, a multi-stream Transformer architecture called SignNet-V2 was developed, which simultaneously recognises sign words and non-manual expression markers from skeletal landmark sequences. The system processes four input modalities, namely body pose, left hand, right hand, and face mesh, through dedicated stream-specific encoders, cross-stream attention fusion, hierarchical temporal encoding, and a multi-task classification head. All models were trained and evaluated under leave-one-signer-out cross-validation, where the test signer never appeared during training, providing a realistic measure of cross-signer generalisation. Under this protocol, SignNet-V2 achieves 38.44% top-1 word recognition accuracy. For expression recognition, a 56-dimensional framework of normalised geometric and temporal features achieves 66.39% overall accuracy under signer-independent evaluation, with neutral and negation recall of 97.22% and 91.18% on the unseen test signer. A second interpretation stage passes recognised signs and expression tags to Google Gemini, which generates grammatically correct Bengali sentences. This work establishes the first published benchmarks for simultaneous word and non-manual marker recognition in BdSL under signer-independent evaluation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Hearing impairment is one of the most widespread but not much discussed issues of public health that is common in the contemporary world. The World Health Organisation reports that more than 1.5 billion individuals all over the earth have some amount of hearing loss, and this number is set to increase exponentially to 2.5 billion by 2050 [1]. One of them, around 430 million people, which is more than 5% of the world's population, need to undergo rehabilitation due to disabling hearing loss. According to the estimates provided by the World Federation of the Deaf, there are about 70 million deaf people in the world, and they use sign language as their main means of communication [2, 3]. The economic cost of unaddressed hearing loss is equally impressive, as it costs the world more than $980 billion every year in annual costs alone [1]. The case is critical in Bangladesh. The second most prevalent disability in the country is hearing loss, with an estimated prevalence of 13.7 million citizens, comprising 9.6 per cent of the country's national population [4, 5]. Out of these, about 1.7 million have severe hearing loss of 90 decibels and above [6]. Today, the number of sign language users in Bangladesh is almost 3 million people, and the technological base to maintain communication is in its infancy and remains in a deplorable state of development and use. Sign language is a visual-spatial natural language that operates on complex gestures, hands, and facial expressions and is used to communicate rather than a spoken language [7, 8]. Sign language has five main parameters, which include handshape, orientation, movement, location, and palm orientation, as its manual components [9, 10]. Importantly, non-manual cues, such as facial expressions, head motions, mouth shapes and body posture, are also included in sign languages and have both emotive and grammatical roles that are required to fully produce a complete communication process [2, 10]. It should be mentioned that sign language has its own grammar and syntax that are not similar to the spoken languages, as well as it does not include the auxiliary words, like is and am [8, 2]. Sign language

is the most natural and necessary communication method by which the Deaf people can share thoughts, feelings, and demands [11, 6]. The Deaf community needs sign language to be the most natural and the most important mode of communication [11, 6]. In addition to a practical use, sign language is a primary element of cultural identity and the foundation of the so-called Deaf Culture [2]. It is based on this importance that in 2000, Bangla Sign Language (BdSL) was given official recognition by the Centre of Disability in Development (CDD) [11].

Existing assistive technology used by the hearing-impaired includes hearing aids, cochlear implants, FM systems, captioning services, and automated Sign Language Recognition (SLR) systems [1, 12]. Nevertheless, the rate of adoption is terribly low; worldwide, only 17% of the individuals who could use a hearing aid actually do so, with the gap being as high as 90% in the African region itself [1]. The problems of SLR technology are multidimensional, including technical challenges such as occlusion, rapid hand movements, and changing lighting conditions [13, 14]. Moreover, certified interpreters are extremely scarce; as of 2015, Bangladesh had only 30 certified sign language interpreters, along with a lack of comprehensive sign language datasets [7, 4]. These technical and resource barriers are compounded by persistent social stigma and communication challenges that limit full access of Deaf people to education, healthcare, and employment [4].

1.2 Rationale of the Study or Motivation

The need to create automated Sign Language Recognition systems arises from the inherent shortage of human interpreter capacity. There is a persistent global deficit of certified sign language interpreters; the United States has 10,000 certified interpreters to serve 48 million Deaf individuals, and Bangladesh had only 30 certified interpreters as of 2015 [7, 4]. This shortage is further aggravated by financial constraints, as the average user can hardly afford to hire professional sign language interpreters even when services are not provided freely, making it an extremely costly endeavour [15, 3]. Digital assistants and communication products remain predominantly voice-oriented and are largely inaccessible to the Deaf community, making sign language the fundamental component of Deaf culture and the primary means of expressing their needs and independence [3, 9]. The consequences of relying on limited manual interpretation are particularly severe in critical domains; the absence of medical personnel proficient in sign language often results in medically underserved populations, which can be life-threatening in emergency situations [16, 7].

There is also a wide communication divide, as hearing-impaired individuals are often rejected or undervalued by those who cannot understand their language, since only a small percentage of the global population is proficient in sign language [8, 9]. These communication disparities have real-world implications: Mistakes in the health care system,

lack of access to health care services, and exclusion from the mainstream school system. Such as using a wheelchair are often unable to meet the needs for inclusivity of disabled students. [4, 12]. When trying to speak to the hearing-impaired individual, this kind of failure can lead to frustration, social withdrawal, and diminished social networks for that particular person. [14, 1]. Automated SLR transformer models has the potential that can help overcome these issues. They can function as interpreters and to translate gestures into written and verbal communication, and they can provide this for deaf people, which can help them interact with the broader community. [17, 15]. In educational context, these robust recognition models can be integrated with self-directed learning and search engines, and they can be easily accessed by the deaf individual even without any human teacher or interaction. [18]. Also in the real world, smart home control is another application which can provide us with virtual reality environments, telemedicine services, and supporting healthcare professionals with accurate and timely diagnosis. [19, 11].

The emphasis on Bangla Deshi Sign Language was driven by a number of compelling reasons. First of all, BDSL is rich in morphology and it has distinct phonology and grammar and a subject-object-verb structure as compared to spoken Bengali and Western language.That alone makes it unique and a very important part of sign language family. [20, 11]. BdSL, as it has a considerable size of alphabet, different expression techniques, lexical indicators, and other features, this complicates the whole process as they often need complex interplay of both. Use hands to communicate ideas and face to convey expressions. [15, 11]. Despite a potential user base of around 13.7 million in Bangladesh, BdSL has low research material. It is an under-researched domain with a severe lack of high-quality benchmark dataset and performance across existing recognition models. [6, 5]. This study has significant social implications. By developing accessible communication tools, it advances the field toward a safer, healthier, and more inclusive society, providing Deaf communities with equal access to social and essential domains [7, 18]. By reducing communication barriers, automated SLR fosters autonomy, confidence, and equitable growth, ultimately minimising the marginalisation of Deaf voices in health policy and design [12, 4]. Furthermore, creating standardised datasets and specialised translation models for regional languages and dialects helps preserve underrepresented languages in the era of artificial intelligence [21, 22].

1.3 Problem Statement

The existing Sign Language Recognition systems, especially those targeting Bangladeshi Sign Language, have a number of inherent weaknesses that hinder their successful implementation and use. Regarding recognition accuracy and performance, most researchers report high performance with self-created datasets, whereas most modern systems face severe challenges on standard benchmark datasets, generally having problems general-

ising to unseen signers [11, 23]. Conventional video-based models use large parameter sizes, including 3D Convolutional Neural Networks, making them time-consuming, computationally intensive, and difficult to deploy on resource-constrained systems, such as smartphones [24, 18]. The current approaches are mostly based on RGB video input, which are sensitive to light and angle variations, has background noises, and these can often lead to misclassification. [25, 14]. The accuracy of recognition tends to decrease for those reasons, and their performance improves significantly with vocabulary size. Systems that can tail well with hundreds of classes have difficulty in managing thousands of signs that are needed for daily communications. [26, 27]. The identification of non-manual elements is extremely lacking as well. Existing methods are very specific in their manual gestures, and they do not take into account the complex linguistic and emotional functions of face expressions and also body postures that are needed to run and complete a proper understanding. Emotional indicators are still poorly understood in SLR research, which leads to misunderstanding and misinterpretation of emotions in legal or medical purposes. [10, 16]. Current systems are only capable of simple labeling, but not natural language or sentences because they are unable to pick up grammatical cues like raised eyebrows or a question or negation or a sense of affirmation.[2, 9]. These are especially hard for BdSL. Bangladesh has a huge user base, over millions of people using it, but it's not enough. BdSL is poorly studied field with lack of profound hearing loss. A few works of systematic research into text-to-gloss translation has been published.[6, 4]. Lack of publicly available information is a serious concern and one of the biggest barriers to BdSL datasets, and another issue is the word level and sentence level recognition, which has insufficient work done on it and requires further progress. [3, 5]. The current systems are not able to record cultural and spatial traditional features of BdSL, which is more compact and localised. Although the signing space is similar to Western sign languages[5]. There is a substantial gap between the size of the spoken Bangla dictionary, approximately 100,000 words, and current sign language resources, approximately 1,200 words, making mapping natural language to sign language extremely challenging [6].

Precisely speaking, neither the beginning nor the end of each gesture is defined. The issue of finding signs in continuous video remains unresolved until then. There are no obvious transitions from one sign to another. [9, 21]. There is a high level of interpersonal variation in signing. speed, style, and physical characteristics results in models that are not robust to signer-independent conditions [28, 23]. There are numerous signs morphologically alike but semantically different. clearly distinguishable, only by subtle spatial relationships between the hand and face, It is hard to resolve existing single-reference models [29]. Furthermore, most of the research takes into account only few aspects of the recognition process; few systems exist that provide a Sign video to fluent natural language (FNL) [30, 7] pipeline. Glosses usually come in sign order (unlike spoken language)

and are current Systems are not so good at reordering and inserting function words in natural. language interpretation [31, 6]. Even the latest models are susceptible to output. errors like omissions, deletions and substitutions, indicating the need for advanced skills. text correction networks and LLM integration to produce understandable results [24]. To overcome these shortcomings, this research will attempt to answer the following questions: (1) How effectively can a multi-stream Transformer architecture with non-manual marker detection enhance BdSL recognition compared to existing baseline models? (2) Can Large Language Models generate natural Bengali sentences from structured sign recognition outputs?

1.4 Objective

The major goal of this study is to create a real-time Bangladeshi Sign Language recog-A new nition and interpretation system which is free from the drawbacks of current systems. It utilizes a combination of deep learning and Large Language Models (LLMs). By using a The model is based on multi-stream architecture (SignNet-V2) and features separate Transformer encoders for the body. This study will attempt to capture both man- and machine-made speech in the pose, left hand, and right hand modalities, as well as face modality. 62 sign classes with ual gestures and critical non-manual features, emotions The grammatical cues (questions, negations), along with the happy, sad, neutral (happy, sad, neutral) are often over-Conventional systems looked at. Moreover, it aims to connect the gap between gloss-By integrating Google Gemini LLM, the system can recognize the level and generate natural language, thereby enhancing the learning experience for the students. To write grammatically and fluently in Bengali. The system works in two modes, conversational and tutoring, to enhance applicability for both real-time commu-For the purposes of information or education. Lastly, the proposed framework is tested against To prove its worth in advancing existing baselines such as BdSL-SPOTER. The BdSL community should have access to accessible communication technologies.

1.5 Methodology in Brief

This study aims to introduce a two-stage hybrid framework of Bangladeshi Sign Language recognition and natural language interpretation, combining specialised deep learning architecture Large Language Model (LLM) enabled tectures. **Stage 1: Sign Recognition.** The The recognition stage is responsible for processing the video stream from the webcam or file source by using MediaPipe Holistic for Skeletal Landmark Extraction. This method allows to retrieve the normalised 3D coordinates for hands, face, and body pose, with robustness against lighting varia- Selected from a vast library of sounds, these samples are used to generate tones and background interference, but are more computationally

efficient than raw. pixel-based methods[8, 24]. A custom processing module is used to process the extracted landmark sequences. It was built using multi-stream architecture in PyTorch, SignNet-V2. The SignNet-V2 uses four body pose, left hand, right hand and Transformer encoders. face modalities that process hand, face and pose modalities independently, allowing spe- Each body region has a cialised feature learning. The results of these four encoders are are aggregated using a Fusion Multi-Layer Perceptron (FusionMLP), which is a composite of the individual ones. High-level representations to a single feature vector. This fused representation feeds into a multi-task classification head (MultiTaskHead) that simultaneously predicts: (i) sign glosses across 62 classes representing the recognised vocabulary, and (ii) non-manual markers across 5 classes (neutral, question, negation, happy, and sad), predicted from a 56 dimensional feature vector of normalised geometric and temporal statistics extracted from facial and head-movement landmarks.

Stage 2: Natural Language Interpretation. The interpretation stage is used for converting the structured recognition result is translated into fluent Bengali sentences. The recognised sign sequence The detected non-manual marker tags and quences are passed to Google Gemini. By using a custom GeminiClient interface, LLM can be transferred via this channel. By thoughtfully prompting the model, When the LLM receives the recognised sign, it will use the detailed instructions provided by system to perform that action. In addition to the detected contextual tag, sequence (e.g., [”Name”, ”What”]). tion”). The output of the LLM is natural and grammatically correct Bengali sentences. For example, when someone asks the question,“আপনার নাম কি?” What's your name? (Your name?). The system Uses context awareness to differentiate between direct conversational mode and educational tutoring mode, powerful safety mechanisms such as retry logic, token. Restrict use of management, and programmed fall-back responses.

Baseline Comparison. The proposed SignNet-V2 is evaluated against BdSL-SPOTER [5], a Transformer-based model that achieved 97.92% top-1 validation accuracy on the BdSLW60 benchmark. Both models are trained and evaluated on the same custom dataset using consistent preprocessing, training epochs, and evaluation metrics to ensure a fair comparison. **Dataset.** This research uses a custom BdSL dataset that includes 62 sign classes representing common vocabulary and 5 non-manual marker categories (neutral, question, negation, happy, sad). The dataset is specifically designed to capture both manual gestures and the critical non-manual components that are frequently overlooked in existing resources.

1.6 Scopes and Challenges

The boundaries of this research are marked by three aspects.The present study is mainly devoted to Bangladeshi Sign Language recognition and interpretation of 62 sign classes

and 5 types of non-manual markers. Tests are carried out in a controlled indoor environment and the system is developed to run in real-time with standard webcam input. The recognition pipeline is designed to support single signer scenarios under the appropriate lighting.

But there are a number of drawbacks to this research. Concerning data scarcity, many sign languages, especially regional ones like BdSL, are low-resource domains and have a scarcity of large-scale training data when compared to spoken languages [31, 30, 11]. The number of publicly available benchmark datasets is very limited, and existing datasets are generally of poor quality, limited in number and diversity and are not suitable for effective model training [11, 14, 21]. The majority of datasets are taken from a small number of native signers and with a small number of repetitions, leading to models that do not generalise to varying signing styles or demographic differences[14, 32]. Creating datasets with gloss annotations is extremely resource-intensive, time-consuming, and expensive, requiring expert signers to manually align video frames with textual representations [31, 21, 24].

Technical challenges involve the requirement of fast and accurate models that can be parallelized, yet be applicable to a wide range of problems.There are also examples of real-time bidirectional communication using the processing[9, 14, 30]. A disadvantage of traditional video-based systems is that they use large-parameter architectures that are computationally inefficient and difficult to deploy on resource-constrained mobile devices [18, 24, 13]. There is high interpersonal variation in signing speed, style, and physical features, which presents significant challenges in making the system robust when the signer is not known [31, 23, 14]. The performance also depends considerably on the scenario, since most models are very sensitive to variations in illumination, background noise, and perspective [14, 25, 13]. Linguistic challenges are also present because sign language is a visual-spatial language with unique grammar and syntax, not merely a gestural reproduction of spoken words [14, 2, 11]. Sign languages vary significantly from one country or region to another, and BdSL has regional dialects and changing signs that pose challenges in the development of a generic recognition system [14, 21, 11]. Continuous SLR is considerably more complex than isolated recognition because the system must detect temporal boundaries in a continuous flow of signs where no clear pauses exist [9, 23, 21]. Practical deployment challenges include hardware and mobile constraints. Many high-performance models are difficult to run on smartphones, often requiring substantial GPU memory and processing power, making them impractical for mobile applications where they are most needed [18, 24, 14]. Systems with low-fidelity avatars or robotic-style movements have generally been rejected by deaf and hard-of-hearing users, as they fail to capture the richness of non-manual signals essential for natural signing [2, 30]. There is also a shortage of human-centred technology capable of directly translating complex visual inputs into fluent natural language within an end-

to-end framework [24, 2]. Specific limitations of this research include a small number of samples per class in the dataset; support for only single-signer recognition without multi-signer capability; the need for adequate lighting to ensure optimal system performance; reduced accuracy for complex sentences with multiple consecutive signs; limited scope for LLM fine-tuning; and incomplete coverage of regional BdSL variations in the current implementation.

Chapter 2

Literature Review

2.1 Sign Language and Deaf Community: Demographics, Communication Barriers, and Accessibility Challenges

Global hearing impairment is a widespread public health problem, with more than 1.5 billion people experiencing some degree of hearing loss worldwide, a figure projected to rise to 2.5 billion by 2050 [1]. Currently, over 430 million people, representing more than 5% of the global population, require rehabilitation for disabling hearing loss [1, 27]. Second most common disability in Bangladesh is hearing loss in particular context. It affects about 13.7 million people or 9.6% of the population [4, 5]. Furthermore, Approximately 1.7 million individuals in Bangladesh have severe hearing loss and Approximately 3000 people use sign language as their native language. [6, 3]. The Deaf and hard-of-hearing (DHH) community is very much impacted by communication. Only a minority of the hearing population is skilled in sign language [8]. This isolation and diminished social networks are a by-product of these barriers for Deaf individuals [14]. The impact is very significant in healthcare, Communication barriers result in misdiagnosis, and restricted access to medical services [4]. Education is also a barrier, and many students with hearing loss face similar education barriers. They make slower progress and drop out of school at higher rates than their hearing peers [1]. Accessibility; A persistent lack of professional support (as of 2015), There were only 30 sign language interpreters in Bangladesh across the country [4]. Traditionally Only 17% of all people with hearing loss who could benefit from using assistive measures do so. [1]. Currently, unaddressed hearing loss has a cost of more than $980 annually for the world. Billion, highlighting the economic importance of scalable and sustainable interventions [1]. Such systemic issues underscore the need for strong, timely and real-time development of solutions. Sign language recognition technology.

2.2 Sign Language Recognition: Evolution and Approaches

The studies for the implementation of Sign Language Recognition (SLR) systems have continued with the application of Sign Language of different paradigms. In fact, prior to 2007, sensor-based solutions were hugely popular, Data sensors such as data gloves, accelerometers, and electromyography were being worn to study. Finger flexion and hand orientation sensors to record hand orientation and finger flexion [23]. These are indeed true, however, methods enabled great accuracy, they were eventually replaced because of major shortcomings, One of the challenges is the high expense of specialized equipment and discomfort of the user [28, 27]. The field was evolving toward the adoption of more natural kinds of interactions, computer vision approaches, and Manufactured features were becoming more prominent based on their handcrafted features. These early vision features from the patterns.The image processing methods were adopted for hand region segmentation and extraction of discriminative features from the patterns for use in systems. Featuring descriptors like Histogram of Oriented Gradients (HOG) and Scale-Invariant. Feature Transformation (SIFT) [19]. The following conventional methods were typically used for classification.Typical classification methods were the following traditional methods. Certain machine learning techniques like Support Vector Machines (SVMs) or Hidden Markov Models (HMMs) [28]. However, these methods that were based on RGB had significant drawbacks because of to be extremely sensitive to background noise, bright lights, etc., and distractibility. illumination [23]. It was in approximately 2015 when the focus changed to deep learning, away from. From handcrafted features to automatic representation learning [28]. Whereas early applications were based on 2D Convolutional Neural Networks for spatial feature extraction, which were soon replaced by 3D CNNs and architectures such as I3D that can process and It is possible to extract spatio-temporal information directly from video volumes [26, 17]. The current state-of-the-art, there is a divergence between appearance-based and skeleton-based modalities. On the other hand, skeleton-based methods have emerged as a promising alternative which relies on pose estimation. Use of libraries such as MediaPipe, OpenPose, etc., to extract structured 2D or 3D keypoints [13]. Recent surveys also indicate that multimodal fusion, obtained by fusion of RGB, skeleton and Provides the most powerful recognition in complex scenarios, graph-based relational data [13]. The research can be divided into a standalone and sequential design. Isolated Sign Language Recognition (ISLR) deals with individual sign words with known. On the other hand, Continuous Sign Language Recognition (CSLR) is to recognize the sign language sequence, while temporal boundaries is to recognize the temporal boundary and find sequences of signs in unsegmented video streams [9]. CSLR is significantly much more difficult as it involves the model to deal with temporal segmentation too and recognition, typically using Connectionist Temporal

Classification (CTC) for sequence alignment[23]. This poses a pressing requirement for Transformer-based models, Employ self-attention to learn the long-range dependencies needed for complete linguistic interpretation [7].

2.3 Challenges in Sign Language Recognition: Non-Manual Markers and the Search for Complete Understanding

Sign language is a natural visual-spatial language which uses the intricate workings of several articulators to convey meaning [1, 27]. Hand movements are the main lexical source in signed communication, however, non-manual markers (NMMs) give the basis of the grammatical and affective meaning of the communicative act [2]. NMMs are behaviors in terms of facial expressions, head movements, body posture and eye gaze according to [14, 10]. These behaviours have a similar function in A.S.L. and B.S.L. as intonation and pitch in spoken languages and provide nuance and context[10, 4]. The real significance of the NMMs is on their capability to change the key sense of a manual sign [16]. For instance, a series of hand gestures may be a statement or a yes/no question, depending on whether the gesture that is raised has raised eyebrows in the signer's face[2]. Furthermore, NMMs serve important grammatical functions, such as signaling denials with head shaking or making conditional statements with special facial expressions [2, 10]. In addition to being important for grammar, they are used primarily as the medium of emotional expression, where the emotions expressed may be of happiness, anger, or sadness [16, 10].

However, in spite of their significance, most of the current SLR systems still struggle to reach a full grasp of the situation as they tend to be primarily manual-based [16]. Facial cues are often treated as a second source of information for many translation systems, resulting in inadequate linguistic interpretation[30]. This hand-centric bias gives rise to productions that do not exhibit the natural co-articulation that occurs in human signing [2, 26]. TToday, NMM recognition is in a state of growth and emergence of multimodal architectures capable of separating these complex signals. For example, the MSKA-SLR framework employs the four-stream attention network to simultaneously process left hand, right hand, face, and body [25]. Even so, there is a significant level of difficulty because of the lack of large datasets with fine-grained NMM annotations and the technical difficulties that come with modelling the style of facial expressions is asynchronous [30, 2].

2.4 Deep Learning and Transformer Models in Sign Language Recognition

In SLR, the spatial feature extraction is done using Convolutional Neural Networks. With respect to BdSL, models like DenseNet201 and ResNet50-V2 have demonstrated themselves well. They give promising accuracy rates of 99% (training) and 93% (testing) on custom datasets [17]. This performance is enhanced further using ensemble approaches; when combining the model Xception with InceptionV3, DenseNet121, ResNet50 and MobileNetV2, the test accuracy of the Ensemble approaches is 99.92% test accuracy on the BDSL-49 dataset. Custom architectures such as KUNet, a shallow CNN optimised using genetic algorithms, reached 99.11% accuracy on the KU-BdSL dataset [15]. Although the 2D CNN works well for spatial mapping, it has limitations for temporal modelling and has led to the introduction of 3D CNN and (2+1)D CNNs [23, 28]. To capture temporal dynamics, Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) [7] had been added. A continuous SLR system using MediaPipe Holistic landmarks paired with a deep LSTM network achieved 88.23% accuracy in real-time Indian Sign Language recognition [8]. The BAUST Lipi study further demonstrated that a CNN-LSTM model reached 97.28% testing accuracy by combining the CNN's spatial feature extraction with the LSTM's temporal modelling capabilities [3]. Despite these successes, RNN-based models are increasingly being surpassed by Transformer-based architectures for modelling long-duration video sequences [31].

Transformers have radically changed the SLR paradigm through the introduction of the self-attention mechanism, which assigns weight to the importance of different input parts regardless of temporal distance [9]. The BdSL-SPOTER framework, a compact four-layer transformer encoder, achieved an unprecedented 97.92% accuracy on the BdSLW60 benchmark [5]. More recently, Rubaiyeat et al. introduced BdSLW401, the largest word-level BdSL dataset to date, comprising 401 sign classes and over 102,000 video samples recorded from 18 native signers in both front and lateral views [33]. The accompanying Relative Quantization Encoding technique anchors landmarks to physiological reference points to reduce signer-specific spatial variability, achieving significant word error rate reductions on multiple benchmarks. Building on this dataset, Shawon et al. explored the fine-tuning of large pre-trained video transformer architectures, including VideoMAE, ViViT, and TimeSformer, for isolated BdSL recognition, achieving 95.5% accuracy on BdSLW60 and 81.04% on BdSLW401 under signer-independent evaluation [34]. Similarly, Mukto et al. developed a MediaPipe-based transformer pipeline for 102 BdSL word classes and achieved 98.1% test accuracy [35]. These recent studies collectively confirm the strong potential of Transformer-based architectures for BdSL word recognition. Crucially, however, none of these works include expression or non-manual

marker labels in their datasets or recognition pipelines, leaving grammatical NMM recognition entirely unaddressed. Transformers are advantageous because they do not require predefined graph structures, and joint relationships can be computed dynamically without explicitly defining a graph [27]. The Sequential Spatio-Temporal Attention Network (SSTAN) applies spatial and temporal multi-head attention to capture both intra-frame and long-range inter-frame dependencies [27]. Graph Convolutional Networks (GCNs) are well-suited for sign language recognition because skeleton data naturally follows a non-Euclidean graph structure [13]. The spatio-temporal dynamic GCN is used in the DSLNet, with an accuracy of 93.70% on WLASL-100[29]. While standard GCNs are limited, however, they can be limited fixed physical graphs which are unable to capture semantic relationships between long-range joints[27]. The performance of a system depends greatly on the input modalities. RGB based models offer detailed visual information, but are prone to noise from background and privacy issues[23]. MediaPipe and OpenPose can extract skeleton-based modalities which are compact, robust, and private [27, 8]. Modern systems increasingly rely on multimodal fusion to combine these complementary features. The GSR-Fusion framework merges RGB, skeleton, and graph-based streams using attention-based fusion, achieving 83.45% accuracy on WLASL [13]. The final area of SLR application is the translation of recognised signs into natural language sentences. Gloss-free translation leverages Large Language Models to avoid resource-intensive gloss annotations [31]. SignBind-LLM employs separate predictors for signing, fingerspelling, and lipreading, and then passes the combined representation to an LLM to generate a sentence [30]. For Bangla specifically, fine-tuning a pretrained mBART-50 on text-to-gloss tasks has shown superior performance, demonstrating that LLMs can effectively handle the shuffling properties of sign language gloss [6].

2.5 Key Findings and Research Gap

Recent developments in SLR have shown that Transformer-based architectures and Graph Convolutional Networks are best suited for capturing the complex spatio-temporal dependencies of signed communication [5, 29]. The highest reported accuracies include a Max Voting Ensemble achieving 99.92% on BdSL alphabets and BdSL-SPOTER reaching 97.92% on the word-level BdSLW60 benchmark [11, 5]. Multi-modal fusion, combining RGB appearance, skeleton landmarks, and graph-based relational data, has been found superior to unimodal methods for handling occlusion and signer variability [13]. Progress in Continuous SLR has been marked by the use of CTC loss for sequence alignment and Transformer-based boundary detection [9, 23].Comprehensible progress in BdSL research has been made, including the publication of datasets like BAUST Lipi and BdSLW60 systematic word-level benchmarks [3].

Despite these achievements, there are still major gaps in their performance. BdSL is

very seriously affected by Despite significant progress in the area of text-to-gloss translation, there are still significant limitations in the availability of comprehensive, sentence-level datasets and a lack of research in this area [21, 6]. One of the main shortcomings is the lack of attention for non-manual markers—most systems only consider the hand gestures and ignore facial expressions, emotion and grammatical markers that are crucial for full linguistic understanding [10, 2, 33, 34, 35]. Moreover, there is a big gap in the natural language output from existing systems as they primarily generate gloss labels and lack tools that account for the specific morphology of Bengali, and to the best of our knowledge, no such example of using LLM for Bengali has been reported [31, 20]. There are no end-to-end pipelines that perform a good job of joining together recognition. In this work, we present an integrated system that combines these three tasks – namely, TST, NMM interpretation, and Natural Language Generation – into a single system. Finally, most of the latest models are still restricted to offline processing and offline mode (conversing without interaction with the tutor; tutoring without simultaneous conversation), and thus not capable of dual mode [24, 18]. The current study aims to fill these gaps by designing a hybrid system consisting of a Transformer-based BdSL recognition system, multi-stream feature extraction of non-manual markers, and a Large Language Model (LLM) for fluent natural language interpretation.

Chapter 3

Methodology

This chapter describes the methodology used to develop a real-time Bangladeshi Sign Language (BdSL) recognition and natural language interpretation framework. The proposed system combines SignNet-V2, a custom multi-stream Transformer model, with a Large Language Model interpretation stage to perform gesture recognition and Bengali sentence generation. The chapter covers system overview, dataset and preprocessing, model architecture, training and evaluation procedures, and impact analysis.

3.1 Final Specifications and Requirements

The system developed in this study is a two-stage pipeline for real-time Bangladeshi Sign Language recognition and natural language interpretation. The functional specifications are defined below.

Input specifications. The system accepts video input from two sources: a live webcam feed for real-time operation, and pre-recorded video files for offline processing. Input frames are processed at the standard capture rate of the input source. The recognition pipeline requires adequate indoor lighting and a single signer positioned within the camera frame.

Stage 1 output specifications. The sign recognition module produces two simultaneous outputs for each input sequence. The first is a predicted sign gloss selected from 62 Bangladeshi Sign Language word classes covering common conversational vocabulary. The second is a predicted non-manual expression label selected from five categories: neutral, happy, sad, negation, and question. Both outputs are produced by a single forward pass through SignNet-V2.

Stage 2 output specifications. The interpretation module accepts the word gloss and expression tag from Stage 1 and produces a grammatically correct Bengali sentence as output. The sentence structure adapts to the expression tag: question markers produce interrogative sentences, and negation markers produce negated constructions.

Functional requirements. The system must operate in two modes. In conversational mode, recognised signs and expression tags are converted to Bengali sentences in sequence, supporting real-time communication. In tutoring mode, the system provides sign-by-sign feedback to support sign language learning. Both modes use the same recognition pipeline and differ only in how the Stage 2 output is presented.

Non-functional requirements. The system runs on a standard consumer computer with a webcam. No specialised sensors or proprietary hardware are required. The software stack is entirely open-source. The Stage 2 module communicates with the Google Gemini API and requires an internet connection for sentence generation. The recognition pipeline includes error handling, retry logic, and fallback responses to maintain operation when API responses are delayed or unexpected.

Constraints. The current implementation supports single-signer scenarios only. Recognition accuracy depends on adequate lighting. The vocabulary is limited to 62 word classes, covering common conversational terms but not the full BdSL lexicon. The system is designed for isolated sign recognition rather than continuous signing streams.

3.2 Risk Management

Several categories of risk were identified during the development of this project and addressed through design and procedural choices.

Dataset risk. The corpus was recorded from seven signers over multiple sessions. The primary risk was inconsistent recording conditions across sessions, which could introduce variation unrelated to signing style. This was mitigated by standardising the recording environment: all sessions used the same indoor setting, fixed camera position, and controlled lighting. A standardised file naming convention was enforced across all recordings to prevent labelling errors during data loading.

Model generalisation risk. Cross-signer generalisation is the core technical challenge for any sign language recognition system evaluated on real users. A model that memorises signer-specific patterns from training data will not work for a new user. This risk was addressed by adopting leave-one-signer-out evaluation from the outset, which forced the training procedure to produce representations that generalise rather than overfit. Early stopping and dropout regularisation were applied to reduce overfitting to the training signers further.

Data quality risk. Landmark extraction can fail on frames where hands or the face are partially occluded or move too quickly for MediaPipe to track. Missing landmarks introduce gaps in the input sequences. This was handled by assigning zero vectors to undetected frames and generating binary attention masks so the Transformer could distinguish valid frames from padding during training and inference.

API dependency risk. The Stage 2 interpretation module depends on the Google

Gemini API for sentence generation. Network failures, API quota limits, and unexpected response formats represent operational risks. These were mitigated through a retry logic for transient failures, token limit management to keep requests within API bounds, response validation to check output language and format, and programmed fallback responses for cases where the API does not return a usable result.

Reproducibility risk. Deep learning experiments can produce different results across runs due to non-deterministic GPU operations. Fixed random seeds were set across PyTorch, NumPy, and Python's random module for all experiments to ensure that reported results are reproducible.

3.3 System Overview

Most sign language recognition systems produce gloss-level outputs without any natural language interpretation, which limits their use in real communication settings [31, 30]. The proposed system addresses this by combining a deep learning recognition stage with a generative language model, forming a two-stage pipeline that converts raw video into fluent Bengali sentences. Figure 3.1 shows the complete workflow.

Stage 1: Sign Recognition. Video input from a webcam or file source is passed to MediaPipe Holistic, which extracts normalised three-dimensional skeletal landmarks for body pose, hands, and face. This skeleton-based representation is robust to lighting changes and background variation, and is more computationally efficient than raw pixel-based approaches [8, 24]. The extracted sequences are then processed by SignNet-V2, a multi-stream Transformer implemented in PyTorch, which uses four dedicated encoders for body pose, left hand, right hand, and face modalities. The fused representation is passed to a classification head that predicts sign glosses across 62 word classes and 5 expression classes simultaneously.

Stage 2: Natural Language Interpretation. The recognised sign sequence and expression tag are passed to Google Gemini via a custom GeminiClient interface. Through prompt engineering, the system instructs the model to generate a grammatically correct Bengali sentence from the recognised gloss sequence and contextual tag. For example, the sequence ["Name", "What"] with a question tag produces "আপনার নাম কি?" (What is your name?). The system operates in two modes, conversational and tutoring, and includes retry logic, token limit management, and fallback responses for reliable operation.

3.4 Dataset and Preprocessing

This section describes the dataset and the preprocessing steps used to convert raw video into normalised landmark sequences for model input.

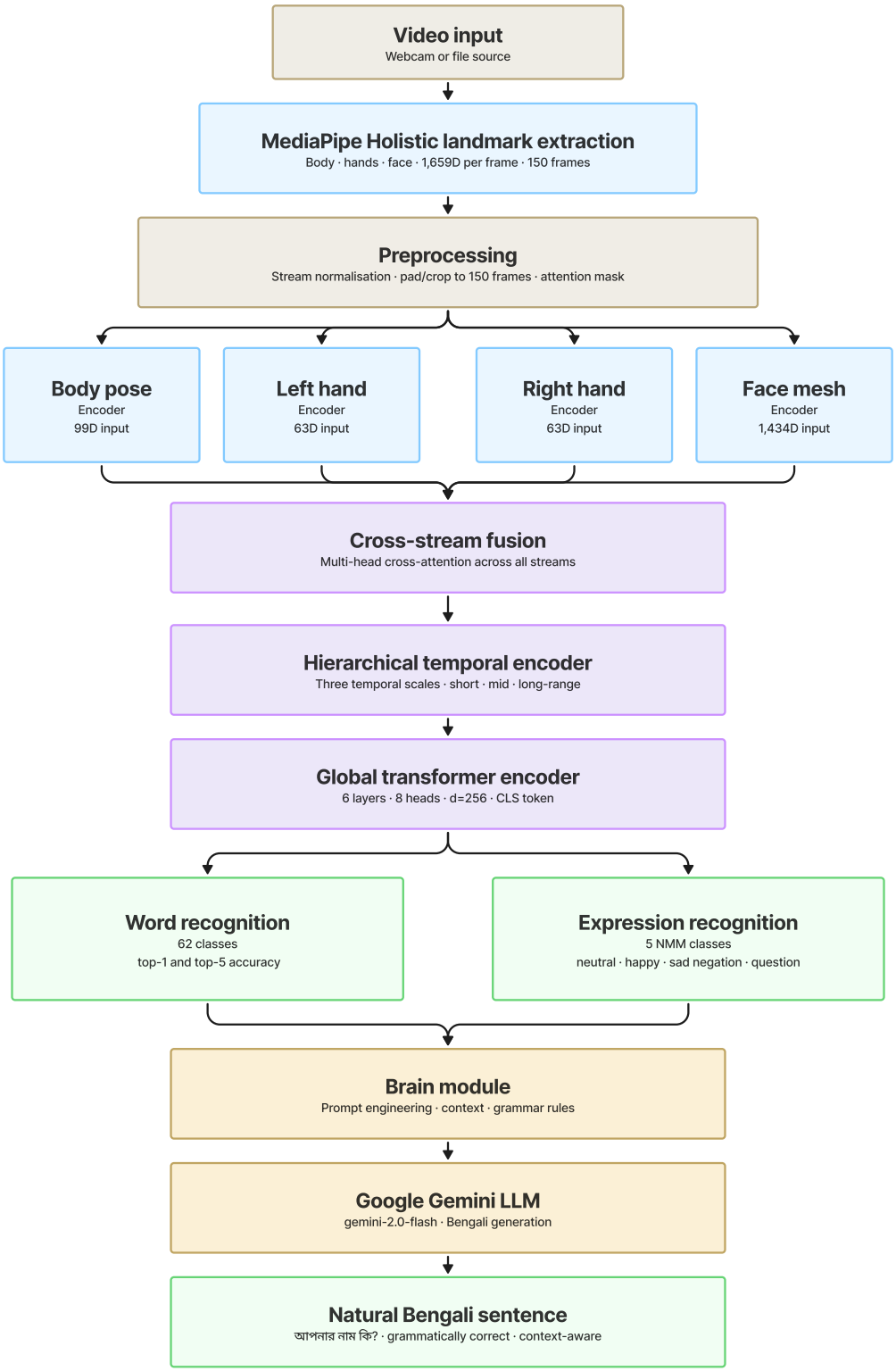

Figure 3.1: System workflow of the two-stage hybrid BdSL recognition and language generation framework.

3.4.1 Dataset Description

The dataset used in this study is BdSL-NMM, an original BdSL corpus recorded specifically to capture both manual and non-manual aspects of Bangladeshi Sign Language. To the best of the authors' knowledge, BdSL-NMM is the first BdSL dataset to include non-manual marker annotations alongside word-level labels. Unlike existing datasets that focus on hand gestures alone, this corpus includes facial expressions and head movements as labelled targets, providing the grammatical and affective context that manual gestures cannot convey independently [10, 2, 16].

The corpus covers 62 sign word classes drawn from everyday conversational Bengali, including pronouns, question words, common action verbs, and emotion-related signs. Expression labels belong to one of five categories: neutral, happy, sad, negation, and question. Negation and question are treated as grammatical markers rather than emotional states, since in BdSL they modify the meaning of entire signed utterances through head shakes and brow raises, respectively [2, 10].

Figure 3.2: The five non-manual expression classes in the BdSL dataset, illustrated using representative frames from Signer S5. From left to right: neutral (no expression), happy (lip corners raised), sad (lip corners lowered), negation (head shake), and question (raised eyebrows with upward head movement). Expressions function as grammatical markers in BdSL, modifying the meaning of the accompanying manual sign.

Seven native signers contributed to the corpus, identified as S1, S2, S3, S4, S5, S6, and S7. All recordings were made in controlled indoor environments under consistent lighting, and each video file follows a standardised naming convention encoding the sign label, signer identity, session index, repetition number, and expression category. This convention allows automatic metadata extraction during data loading without manual annotation at runtime. The vocabulary selection and expression category design were validated by Mohammad Arish Abedin, Life Member of the Bangladesh National Federation of the Deaf and Vice President of the Bangladesh Student Welfare Federation of the Deaf. As a native BdSL user, he confirmed that the selected sign words are widely used within the Deaf community and appropriate in the BdSL context, and that the five expression categories are acceptable as non-manual markers in BdSL.

The dataset is partitioned using a leave-one-signer-out protocol. Signer S7 is held out entirely as the test set and does not appear in training or validation at any point. The remaining six signers form the training and validation pool, split approximately 90/10 within that group. Table 3.1 presents the full dataset statistics under the primary

leave-one-signer-out protocol.

Table 3.1: Dataset statistics under the leave-one-signer-out evaluation protocol.

Split	Samples	Percentage	Word Classes	Signers
Training	3,214	77.0%	62	6 (S1, S2, S3, S4, S5, S6)
Validation	357	8.6%	62	6
Test	601	14.4%	62	1 (S7 only)
Total	**4,170**	**100%**	**62**	**7**

The test signer S7 contributes 601 samples covering all 62 word classes and all 5 expression categories. Because S7 never appears during training, evaluation on this split measures genuine cross-signer generalisation rather than within-signer generalisation memorisation [28, 23]. Table 3.2 presents the per-signer breakdown of the corpus, showing the number of files, word class coverage, and expression distribution for each signer.

Table 3.2: Per-signer file counts and expression class distribution in BdSL-NMM. N = neutral, H = happy, Sa = sad, Ne = negation, Q = question.

Signer	Files	Words	N	H	Sa	Ne	Q
S1	594	60	140	119	118	100	117
S2	544	59	109	111	110	88	126
S3	634	60	125	128	127	126	128
S4	598	60	120	119	120	119	120
S5	599	60	141	119	119	102	118
S6	600	60	142	119	120	101	118
S7	601	60	141	119	120	102	119
Total	**4,170**	**62**	**918**	**834**	**834**	**738**	**846**

Six of the seven signers covered 60 of the 62 word classes individually. Signer S2 covered 59 classes. Full coverage of all 62 unique word classes is achieved across the combined corpus. Expression labels are near-balanced across classes, with each category accounting for between 17% and 22% of the total 4,170 files. The minimum number of samples for any single word class is 22, with a mean of 67.26 and a median of 69 files per class.

3.4.2 Feature Extraction and Normalisation

Raw video frames are converted into normalised landmark sequences through four sequential steps: landmark extraction, stream-specific normalisation, sequence processing, and data augmentation.

Landmark Extraction. MediaPipe Holistic extracts skeletal landmarks from each frame, producing four distinct sets: body pose (33 landmarks), left hand (21 landmarks), right hand (21 landmarks), and face mesh (478 landmarks, using `refine_landmarks=True`).

Each landmark carries three-dimensional coordinates (x, y, z), giving feature dimensions of 99 for body pose, 63 for each hand, and 1,434 for the face. The complete per-frame representation is 1,659-dimensional. Frames where landmarks are not detected due to occlusion or rapid movement are assigned zero vectors to maintain temporal alignment.

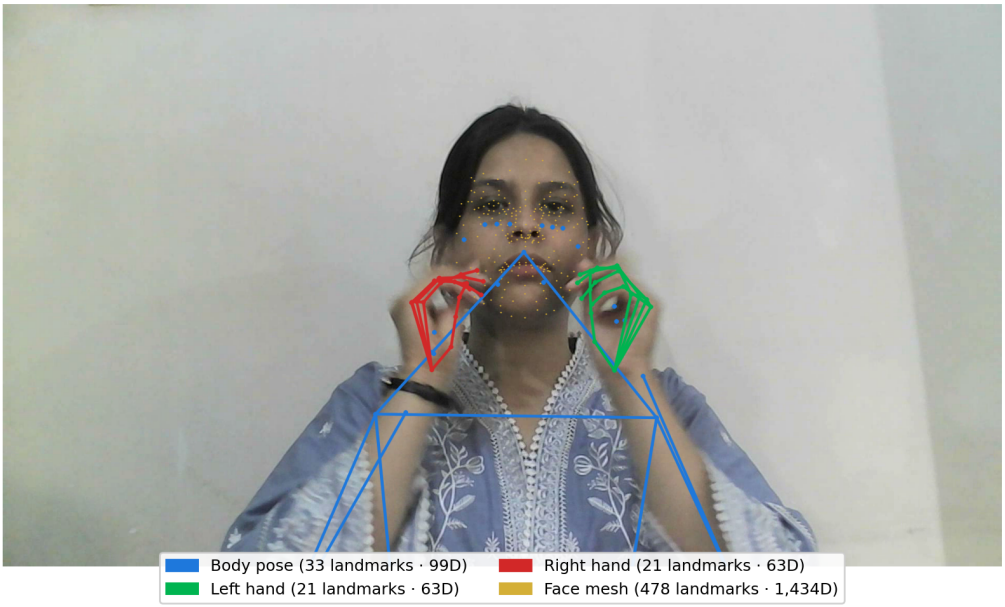

Figure 3.3: MediaPipe Holistic landmark extraction applied to a sample frame from Signer S5, showing all four input streams: body pose (blue, 33 landmarks, 99D), right hand (red, 21 landmarks, 63D), left hand (green, 21 landmarks, 63D), and face mesh (gold, 478 landmarks, 1,434D). Landmarks are labelled from the signer's perspective. The complete per-frame representation is 1,659-dimensional.

Normalisation. Stream-specific normalisation is applied to remove variation caused by differences in signer body size, hand scale, and camera distance. Body pose landmarks are normalised relative to the shoulder midpoint, with shoulder width as the scale factor. Hand landmarks are normalised relative to the wrist, scaled by palm spread. Face landmarks are centred at the face centroid and scaled by the bounding box of the face region. The same equation applies across all streams:

$$\mathbf{p}_{norm} = \frac{\mathbf{p} - \mathbf{c}_{ref}}{s_{ref}} \tag{3.1}$$

where $\mathbf{p}$ is the original landmark coordinate vector, $\mathbf{c}_{ref}$ is the stream-specific reference centre, and s_{ref} is the anatomically derived scale factor.

Sequence Processing. All sequences are padded or centre-cropped to a fixed length of 150 frames for batch processing. Short sequences receive zero padding at the end. Long sequences are centre-cropped to retain the most informative portion of the sign. Binary attention masks distinguish valid frames from padding positions during Transformer attention computation.

Data Augmentation. Several augmentation strategies are applied during training to reduce overfitting given the limited dataset size. Temporal scaling randomly adjusts sequence speed between 0.8× and 1.2× using linear interpolation. Spatial augmentations include Gaussian noise with standard deviation $\sigma = 0.02$, two-dimensional rotation within $\pm 15°$, and uniform scaling between 0.9× and 1.1×. Mixup augmentation with $\alpha = 0.2$ creates interpolated training samples by linearly combining pairs of input sequences and their labels, which reduces overfitting in gesture recognition tasks [13].

3.5 Model Architecture

SignNet-V2 processes each input modality through a dedicated encoder before fusing the streams, allowing the model to learn body-region-specific representations while capturing cross-modal relationships [25]. Figure 3.4 shows the full architecture.

The architecture has five components: multi-stream input encoders, cross-stream fusion, hierarchical temporal encoder, global Transformer encoder, and classification head.

Multi-Stream Input Encoders. Each of the four streams passes through a separate StreamSpecificEncoder. Input features are projected to a shared embedding dimension d_{model} using a linear layer, followed by layer normalisation and dropout. Temporal convolution blocks with residual connections model short-range dependencies using 1D convolutions with kernel size 3. Spatial attention computes relationships among landmarks within each frame, and positional encodings preserve temporal order information [9].

Cross-Stream Fusion. The CrossStreamFusion module exchanges information across streams using multi-head cross-attention. Stream features are concatenated and projected to form a shared key-value representation. Each stream then queries this representation through separate attention heads. The attended outputs are averaged and combined with a residual connection from the original stream, retaining stream-specific information while incorporating context from other modalities.

Architectural Design Rationale. The decision to process each body region through a dedicated encoder, rather than concatenating all landmarks into a single flat input, reflects a key observation about sign language structure. Hand gestures, body posture, and facial expressions operate on different spatial and temporal scales. A hand sign may change within a few frames, while a grammatical head movement unfolds over a longer duration. A single shared encoder applied to a concatenated 1,659-dimensional input would need to learn these different dynamics simultaneously from the same weights, which places conflicting demands on the model. Separate stream encoders allow each modality to develop representations suited to its own motion characteristics before the streams are merged. Cross-stream fusion then combines these specialised representations, capturing interactions between modalities, such as how a head shake during a particular

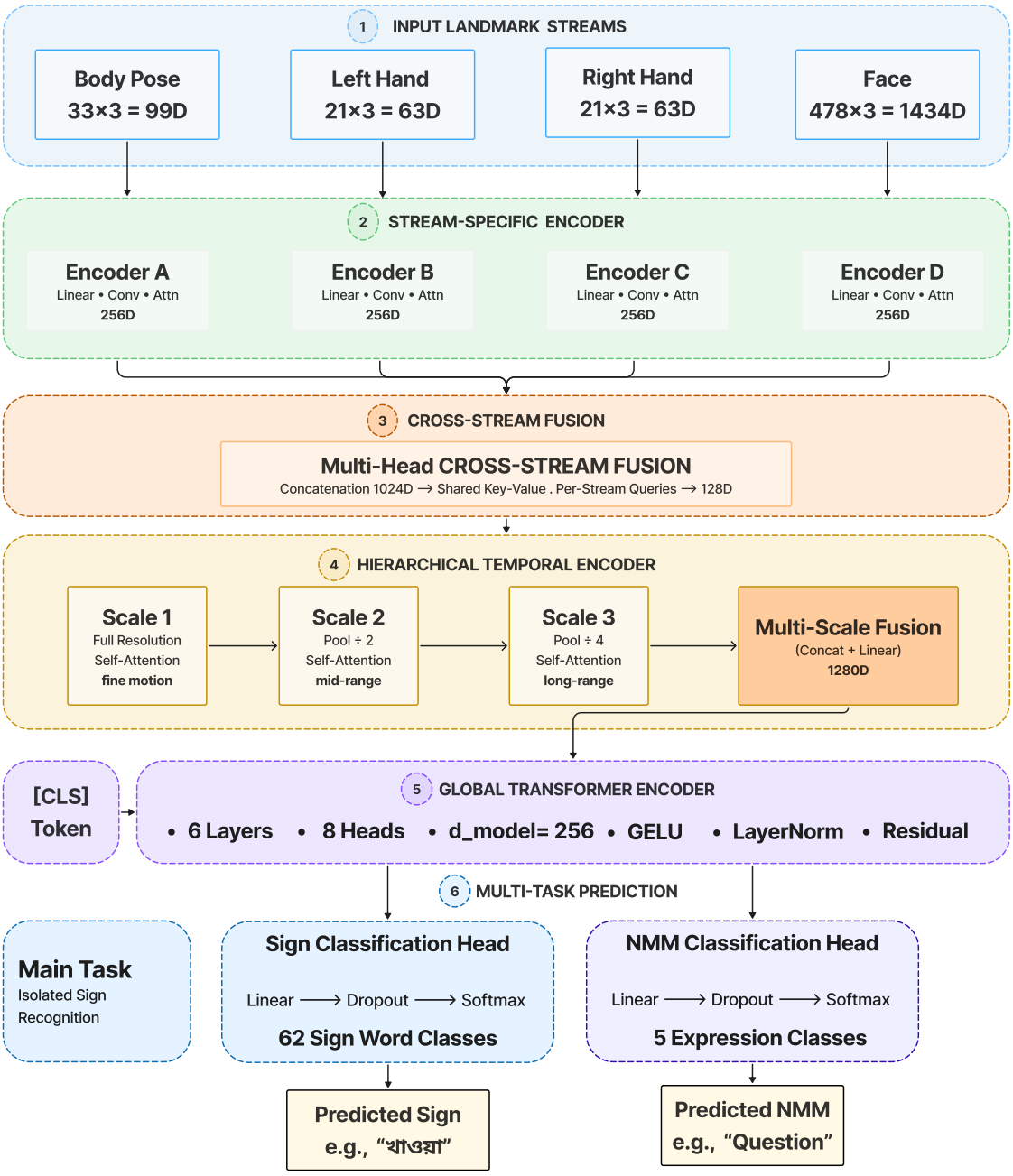

Figure 3.4: SignNet-V2 architecture with multi-stream encoders for pose, hands, and face, followed by cross-stream fusion, hierarchical temporal encoding, global Transformer processing, and a multi-task classification head predicting 62 word classes and 5 expression classes.

hand configuration signals negation, rather than treating the two signals as independent features. This design follows the principle established in multi-stream attention architectures for sign language recognition, where stream separation before fusion consistently outperforms flat concatenation [25, 27].

Hierarchical Temporal Encoder. The HierarchicalTemporalEncoder captures temporal patterns at three scales using average pooling with kernel sizes 1, 2, and 4. Multi-head self-attention is applied at each scale independently. The multi-scale outputs are interpolated back to the original sequence length, concatenated, and fused through a projection layer with layer normalisation. This allows the model to represent both fine-grained finger articulations and slower phrase-level transitions [27].

Global Transformer Encoder. The GlobalTemporalEncoder applies multiple Transformer layers with pre-layer normalisation [9]. Each layer uses 8-head self-attention followed by a position-wise feed-forward network with GELU activation. A learnable class token is prepended to the sequence and aggregates global information for classification.

Classification Head. The class token representation passes through a three-layer MLP with GELU activations and dropout between layers. The first layer preserves the embedding dimension, the second halves it, and the output layer produces logits over 62 sign classes. Layer normalisation is applied before the first linear layer. The expression branch uses a 56-dimensional feature vector of normalised geometric and temporal statistics derived from seven facial and head-movement signals, described in full in Section 4.3.

LLM Integration. Recognised glosses and expression tags are passed to Google Gemini through a custom GeminiClient interface. The prompt engineering strategy supplies the LLM with the sign sequence, the detected grammatical marker, and instructions for generating a Bengali sentence following natural grammar patterns. The interface includes retry logic, token limit control, response validation, and fallback responses.

3.6 Training and Evaluation

This section describes the training procedure and evaluation approach used across all experiments.

Leave-One-Signer-Out Protocol. All primary experiments follow a leave-one-signer-out (LOSO) evaluation protocol. Signer S7 is excluded from training and validation entirely; the model is trained on the remaining six signers and evaluated on S7's 601 samples. This protocol tests whether the model generalises to a person it has never seen, which is the relevant capability for real-world deployment [28, 23]. It is considerably stricter than random-split evaluation, where samples from the same signer appear in both training and test sets.

3.6.1 Training Configuration

The model is optimised using AdamW with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.95$ and decoupled weight decay. A OneCycleLR scheduler applies cosine annealing with 10% linear warmup, which stabilises early training and supports convergence. Cross-entropy loss with label smoothing encourages calibrated probability estimates and reduces overconfidence on training classes.

Automatic mixed precision (AMP) training performs forward passes in float16 while accumulating gradients in float32, reducing memory consumption and accelerating computation on GPU hardware. Gradient clipping with a maximum norm threshold prevents instability from large gradients in early epochs. Early stopping halts training when validation performance does not improve for a fixed number of epochs. All experiments were trained for a maximum of 50 epochs, with early stopping applied in practice to halt training when validation performance plateaued. All experiments use fixed random seeds for reproducibility and are trained on an NVIDIA RTX 4080 Super GPU. Table 3.3 lists the hyperparameters for both baseline and optimised configurations.

Table 3.3: Training configuration and hyperparameters for SignNet-V2.

Parameter	**Baseline**	**Optimised**
Embedding dimension (d_{model})	128	256
Transformer layers	4	6
Attention heads	8	8
Feed-forward dimension	512	1024
Dropout rate	0.2	0.3
Learning rate	3×10^{-4}	1×10^{-4}
Weight decay	0.05	0.1
Label smoothing	0.1	0.2
Mixup alpha	0.2	0.4
Batch size	16	12
Epochs	50	50
Gradient clip norm	1.0	0.5
Early stopping patience	25	15

3.6.2 Evaluation Metrics

Top-k accuracy for $k \in \{1, 5\}$ measures the proportion of test samples where the correct class appears within the top k predictions. This captures both precise recognition (top-1) and the model's ability to rank the correct sign highly even when the single best prediction is wrong (top-5). Precision, recall, and F1-score are computed with macro-averaging, giving equal weight to all 62 classes regardless of sample frequency. A 62×62 confusion matrix identifies systematic misclassification patterns between visually similar

signs. For expression recognition, per-class recall and overall accuracy, macro F1, and macro recall are reported across the five expression categories.

3.7 Societal Impact

Bangladesh has approximately 1.7 million people with severe hearing loss and around 3 million sign language users, yet certified BdSL interpreters remain scarce [4, 6]. The communication gap this creates is not simply inconvenient; it affects access to healthcare, education, and employment in concrete ways. Patients who cannot communicate with medical staff face misdiagnosis. Students in mainstream classrooms without interpreter support fall behind. Job applicants who rely on sign language are routinely excluded from workplaces that have no mechanism to bridge the language gap [4, 12].

The system developed in this study addresses part of this problem by providing an automated recognition pipeline that requires no specialist interpreter. A user can sign in front of a standard webcam and receive a Bengali sentence as output, making the tool accessible on consumer hardware without additional cost. The two operational modes, conversational and tutoring, extend the application to both everyday communication and sign language education. In a tutoring context, a learner can practise signs and receive immediate feedback, reducing dependence on human instructors who are in short supply.

The inclusion of grammatical non-manual markers, specifically negation and question, has direct social value. A system that recognises only hand gestures produces ambiguous output; the same hand sign means different things depending on whether the signer shakes their head or raises their eyebrows. Capturing these markers produces interpretations that more accurately reflect what the signer intended, which matters in any context where miscommunication carries real consequences.

The skeleton-based input representation also addresses a concern that is sometimes overlooked in accessibility technology: privacy. Because the system processes landmark coordinates rather than raw video frames, it does not store or transmit identifiable visual images of the user. Sample video frames are included in this document for illustrative purposes only, with the full informed consent of all signers. This makes the pipeline more suitable for deployment in sensitive settings such as medical consultations or legal proceedings, where video recording of individuals without consent raises ethical and legal concerns [14].

Finally, developing a purpose-built dataset and recognition model for BdSL contributes to the longer-term goal of preserving and documenting a language that remains under-resourced in the artificial intelligence literature [21]. Every dataset and benchmark created for a low-resource sign language reduces the barrier for future researchers working on the same problem.

3.8 Environmental Impact

The environmental footprint of this project is modest relative to comparable deep learning research. The system relies entirely on software and standard consumer hardware; no specialised sensors, data gloves, or proprietary devices were used or manufactured for this study. MediaPipe Holistic runs on CPU for landmark extraction, and model training was performed on a single NVIDIA RTX 4080 Super GPU rather than a large distributed cluster.

Training runs were kept to a maximum of 50 epochs per configuration, and early stopping further reduced unnecessary computation in practice. The use of mixed precision training halved the memory footprint per forward pass, which reduced the energy draw per training step. Compared to video-based approaches that process raw RGB frames through large 3D convolutional networks, the skeleton-based pipeline used here is substantially more energy-efficient, since it operates on compact coordinate vectors rather than high-dimensional pixel arrays [24].

The trained models are lightweight, with SignNet-V2 containing approximately 10.3 million parameters. Inference runs in real time on a single GPU without requiring cloud compute, which means deployment does not incur the ongoing energy cost associated with server-based cloud inference at scale.

No physical materials were consumed beyond standard electricity use. Dataset collection involved recording video with existing cameras in an indoor room; no travel, shipping, or hardware procurement contributed to the project's carbon footprint. The dataset itself is stored as compressed NumPy arrays, keeping storage requirements minimal.

3.9 Ethical Considerations

Informed Consent and Signer Privacy. All participants gave informed consent for the use of their recordings in this thesis, in any associated publications, and for illustrative academic purposes. Sample frames are included in this chapter for methodological illustration. As described in Section 3.4.1, the system processes skeletal landmark coordinates instead of video, so no identifiable visual representation of any of the signers is used during training or inference. The vocabulary and expressions of the dataset are presented in the table below. The vocabulary and expression categories were also checked with a certified Deaf representative, who acknowledged their linguistic suitability and gave consent for his name, image, and designation to be referenced in this work.

Dataset Bias and Representativeness. The dataset includes seven signers and is a small sample of the total signing population in BdSL. Regional variation, age variation, and signing experience all have an impact on the ways in which people produce signs, and this diversity cannot be adequately captured in a seven-signer corpus. This should

be taken into account when interpreting the results discussed in Chapter 5. The leave-one-signer-out evaluation protocol explicitly addresses this by testing the model on a signer not encountered during training, which provides a more realistic assessment of generalisation than a random-split evaluation would provide.

Risk of Misrecognition. No automatic recognition system is foolproof, and every system available on the market has its own limitations. Any recognition error in a communication aid has consequences. An incorrect word prediction, a different interpretation of elements in the same sentence, or the absence of a negation marker may change the meaning of a signed utterance. This can make the output confusing and sometimes harmful, especially in medical or legal contexts. As implemented now, this is a research prototype and is not certified for use in high-stakes environments. Any future deployment should involve evaluation by representative members of the Deaf community and should include mechanisms for users to check and correct the output of the system.

LLM Output Responsibility. The Stage 2 interpretation module uses Google Gemini to generate Bengali sentences. LLM outputs can be incorrect, biased, or culturally inappropriate. The system has response validation and fallback mechanisms. However, these are not a guarantee of correctness. All users of the deployed system should be informed that the generated sentences are computer-generated and may require verification or adjustment.

Academic Integrity. All code, models, and experimental results presented in this thesis are original work created by the research team. External models, specifically the published work in [5] , were reproduced as BdSL-SPOTER for comparison under the same dataset and evaluation conditions, and are clearly identified as reproduced baselines throughout.

3.10 Project Management Plan

The project was carried out over approximately eight months, from dataset planning through final evaluation and thesis writing. Table 3.4 summarises the major phases and their approximate durations.

Team responsibilities were divided according to expertise. Dataset collection and annotation were a shared team effort. Model architecture design and training were led by two members with primary responsibility for the deep learning pipeline. LLM integration and the Brain module were handled separately by one team member. Thesis writing, result analysis, and documentation were distributed across the full team with a designated lead for each chapter.

Regular progress meetings were held weekly with the supervisor to review experimental results, resolve technical issues, and adjust the development timeline as needed. Version control was maintained throughout using Git, with all training scripts, model

Table 3.4: Project timeline and major milestones.

Phase	Activities	Duration
Dataset Design	Sign selection, naming convention, recording protocol	3 weeks
Data Collection	Recording sessions with 7 signers	4 weeks
Preprocessing	Landmark extraction, normalisation, NPZ conversion	3 weeks
Baseline Training	SingleStreamBaseline and BdSL-SPOTER reproduction	3 weeks
Model Development	SignNet-V2 architecture, expression head, investigation	6 weeks
Model Optimisation	Padding mask fix, LOSO evaluation	3 weeks
LLM Integration	Brain module, GeminiClient, prompt engineering	2 weeks
Thesis Writing	All chapters, figures, final review	4 weeks

checkpoints, and evaluation results stored in a shared repository.

3.11 Economic Analysis

The most important economic reason for this work is the cost of human sign language interpretation. Professional interpreters in Bangladesh are very few, and hiring them for extended periods can be very expensive for individuals or institutions [4, 15]. n situations such as hospital visits or legal advice, the lack of accurate communication can create serious problems. The absence of an interpreter is not just an inconvenience; it means that decision-making may take place without accurate communication.

This study develops the system using a standard consumer computer and webcam. In addition to the hardware cost, most institutions already use similar equipment for other software applications. The software stack is open-source: Python, PyTorch, and MediaPipe do not require any licensing fees. The Google Gemini API used for Stage 2 interpretation is available at affordable prices for medium consumption levels. There are no service charges, no major running expenses, no need for trained specialist staff to operate the system, and no requirement for long-term service contracts.

The tutoring mode is an affordable addition to education institutions. A student can practice sign vocabulary independently and receive immediate recognition feedback, reducing the number of supervised sessions needed under the guidance of a human teacher. This decreases the cost of sign language education per student over time, especially where instructor availability is limited.

The system is not designed to replace human interpreters in all situations, particularly high-stakes ones. It is an inexpensive tool that can improve access to basic communication and interpretation support when no interpreter is available.

Chapter 4

Design and Implementation

This chapter outlines the decisions made in the design of the recognition system. The methodology and architecture are discussed in Chapter 3. The focus here is on why certain configurations were selected, the relationships between the three model designs, and how the expression recognition framework was specified. Implementation details are provided in the following chapters.

4.1 Design Overview

The main design challenge in this work was the development of a sign language recognition system that could generalise to a signer it had never seen before. Most current BdSL systems report accuracy on data that is similar to the data used for training, which may inflate reported accuracy. Since the system was designed from the beginning for leave-one-signer-out evaluation, this influenced all design decisions, from separating input streams to selecting the final model.

The initial architectural choice was to send each part of the body through its own encoder instead of merging all landmarks into a single flat vector. The landmarks are grouped into separate streams rather than concatenated into one flat vector. Hand movements, body posture, and facial signals have distinct temporal patterns and carry different types of information. Treating them the same through a shared encoder may discard this structure. Separate encoders allow each stream to develop representations appropriate to its modality before the streams interact through cross-attention fusion. The rationale for this design is covered in Section 3.3. What follows here is how that design was operationalised across three configurations, each serving a distinct purpose.

A second design decision concerned expression recognition. Non-manual markers in BdSL, specifically negation and question, are grammatical rather than purely affective. They change the meaning of an entire utterance. Treating them as a secondary output of the same model that recognises words, rather than as a separate post-processing step,

was a deliberate choice to keep the two tasks coupled during training. The tradeoffs of that choice are examined in Chapter 5.

4.2 Model Configurations as Design Alternatives

Three model configurations were designed and trained, each addressing a specific question about the recognition problem. Table 4.1 summarises the configurations.

Table 4.1: Summary of model configurations and their design purpose.

Model	Streams	Design Purpose	Expression
SingleStreamBaseline	1 (all concatenated)	Establish task difficulty under LOSO evaluation	No
BdSL-SPOTER (reprod.)	1 (all concatenated)	External published reference point under identical conditions	No
SignNet-V2	4 (separate)	Proposed multi-stream architecture with simultaneous word and expression recognition	Yes

4.2.1 SingleStreamBaseline

The SingleStreamBaseline was designed to answer a basic question: how well does a standard transformer perform on this task when given all the landmark data at once? All 1,659 features per frame are concatenated into a single vector and passed through a four-layer transformer encoder with a word classification head. There is no stream separation, no cross-attention, and no expression output. This configuration was not designed to be competitive. Its purpose is to establish a lower-bound reference for what a reasonable but unspecialised design achieves under leave-one-signer-out evaluation. Any proposed model that does not clearly exceed this baseline cannot justify its additional complexity.

4.2.2 BdSL-SPOTER (Reproduced)

BdSL-SPOTER [5] is a published transformer-based model that reported 97.92% accuracy on the BdSLW60 benchmark under a random split protocol. To use it as a meaningful reference, it was reproduced and evaluated under the same leave-one-signer-out protocol and dataset used in this study. The input was adapted to accept the full 1,659-dimensional landmark vector through a linear projection layer, keeping the original six-layer transformer architecture and sinusoidal positional encoding intact. No expression head was

added. This configuration provides a point of comparison with published work under controlled and equivalent conditions, which a direct citation of the original accuracy figure cannot.

4.2.3 SignNet-V2

SignNet-V2 is the proposed architecture, trained from random initialisation on the LOSO splits. It uses all four stream-specific encoders, cross-stream fusion, hierarchical temporal encoding, and a global Transformer encoder. The classification head has two branches: a word classifier predicting across 62 sign classes, and an expression classifier predicting across 5 non-manual marker classes. Both branches are optimised simultaneously throughout training using a combined loss over word and expression targets.

This multitask design was a deliberate choice. Negation and question markers in BdSL change the grammatical meaning of an entire utterance, not just its emotional tone. Keeping word and expression recognition coupled in a single model means the shared encoder learns representations that carry both lexical and grammatical information, rather than treating the two tasks as independent problems. The cost of that choice, in terms of word accuracy relative to a word-only model, is examined in Chapter 5.

Training from scratch on the LOSO data tests whether the multi-stream design, without any prior knowledge of signing styles, produces generalisable representations for an unseen signer. This configuration is the primary proposed model evaluated throughout Chapter 5.

4.3 Expression Recognition Design

Designing the expression recognition component required specifying which physical signals correspond to each expression class and how to extract those signals in a way that generalises across signers with different facial geometry.

4.3.1 Signal Selection and Feature Extraction

The five expression classes fall into two categories. Negation and question are grammatical markers realised through head movement: negation through a lateral head shake and question through a vertical nod combined with eyebrow raising. Neutral, happy, and sad are affective states realised primarily through lip and mouth configuration. This distinction guided signal design: signals for grammatical markers were derived from head-position landmarks, while signals for affective states were derived from face mesh landmarks around the mouth region.

Seven normalised signals are computed per frame. Let N denote the nose landmark, E_L, E_R the left and right ear landmarks, L_L, L_R the lip corners, L_U, L_D the upper and

lower lip, B_L, B_R the eyebrow landmarks, I_L, I_R the eye landmarks, and F_L, F_R the lateral face-boundary landmarks. The ear midpoint $E_M = (E_L + E_R)/2$, ear distance $D_{ear} = \|E_L - E_R\|_2$, and face width $D_{face} = \|F_L - F_R\|_2$ serve as normalisation denominators, making all measurements invariant to signer size and camera distance.

- **HeadYaw**: horizontal nose displacement relative to the ear midpoint, scaled by ear distance. Primary signal for negation.

$$\text{HeadYaw} = \frac{N_x - E_{M,x}}{D_{ear}} \tag{4.1}$$

- **HeadNod**: vertical nose displacement relative to the ear midpoint, scaled by ear distance. Primary signal for question markers.

$$\text{HeadNod} = \frac{N_y - E_{M,y}}{D_{ear}} \tag{4.2}$$

- **HeadRoll**: vertical ear-to-ear difference scaled by ear distance, distinguishing head orientation from horizontal movement.

$$\text{HeadRoll} = \frac{E_{L,y} - E_{R,y}}{D_{ear}} \tag{4.3}$$

- **EyebrowEye**: mean vertical distance between eyebrow and eye landmarks, scaled by face width. Captures eyebrow raising for question markers.

$$\text{EyebrowEye} = \frac{\frac{1}{2}[(B_{L,y} - I_{L,y}) + (B_{R,y} - I_{R,y})]}{D_{face}} \tag{4.4}$$

- **MouthWidth**: lip-corner span scaled by face width. Smiling increases this value.

$$\text{MouthWidth} = \frac{\|L_L - L_R\|_2}{D_{face}} \tag{4.5}$$

- **MouthOpen**: vertical lip aperture scaled by face width.

$$\text{MouthOpen} = \frac{\|L_U - L_D\|_2}{D_{face}} \tag{4.6}$$

- **LipCornerHeight**: vertical offset of the average lip corner from the mouth centre, scaled by face width. Positive values correspond to raised corners (happy), negative to lowered corners (sad).

$$\text{LipCornerHeight} = \frac{\frac{1}{2}(L_{L,y} + L_{R,y}) - \frac{1}{2}(L_{U,y} + L_{D,y})}{D_{face}} \tag{4.7}$$

4.3.2 Temporal Aggregation

For each of the seven signals $s = [s_1, \ldots, s_T]$, eight sequence-level statistics are computed: mean, standard deviation, minimum, maximum, range, mean absolute velocity, maximum absolute velocity, and end-to-start change. Formally:

$$\mu_s = \frac{1}{T} \sum_{t=1}^{T} s_t \tag{4.8}$$

$$\sigma_s = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (s_t - \mu_s)^2} \tag{4.9}$$

$$\text{Range}_s = \max_t(s_t) - \min_t(s_t) \tag{4.10}$$

$$\text{MeanVel}_s = \frac{1}{T-1} \sum_{t=2}^{T} |s_t - s_{t-1}| \tag{4.11}$$

$$\text{MaxVel}_s = \max_{t=2,\ldots,T} |s_t - s_{t-1}| \tag{4.12}$$

$$\text{Change}_s = s_T - s_1 \tag{4.13}$$

This yields $7 \times 8 = 56$ features per recording. Static features such as mean mouth width capture sustained facial expressions, while dynamic features such as range and velocity capture head-motion patterns. The primary distinguishing signals per class are as follows: neutral relies on low head movement and typical mouth geometry; happy on increased mouth width and upward lip-corner height; sad on downward lip-corner position and altered mouth shape; negation on large HeadYaw range and velocity; and question on HeadNod range, direction of change, and elevated EyebrowEye.

4.3.3 Classifier Selection

Three classifiers were evaluated on the 56-dimensional feature vectors: standardised multinomial logistic regression, a standardised two-layer multilayer perceptron, and a random forest with 500 trees trained with class-balanced sampling. Model selection used validation macro F1, giving equal weight to all five expression classes. The random forest produced the best results and was used for all reported evaluations. Results for both random-split and leave-one-signer-out protocols are reported in Section 5.1.

4.4 Implementation Details

All models were implemented in PyTorch and trained on a single NVIDIA RTX 4080 Super GPU with 16GB of VRAM. Mixed precision training using PyTorch's automatic

mixed precision package was enabled throughout to reduce memory consumption and accelerate forward passes without affecting numerical precision of gradient accumulation.

Training runs were managed and monitored using Weights and Biases, which logged loss curves, validation accuracy, learning rate schedules, and gradient norms at each epoch. Fixed random seeds were set across PyTorch, NumPy, and Python's random module to ensure reproducibility across runs.

The full software stack used in this project is listed below. All tools and libraries are open-source and freely available.

- **Python 3.10**: primary programming language
- **PyTorch 2.1**: model implementation and training
- **MediaPipe 0.10**: landmark extraction from video frames
- **NumPy / SciPy**: numerical processing and feature computation
- **Weights and Biases**: experiment tracking and visualisation
- **Google Gemini API**: natural language generation in Stage 2

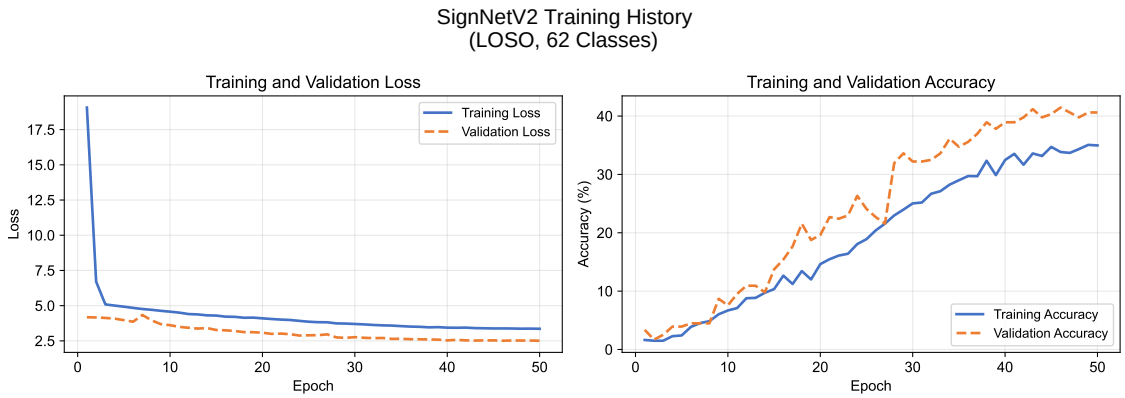

Figure 4.1: Training and validation loss and accuracy curves for SignNet-V2 under leave-one-signer-out training on BdSL-NMM (62 classes). Validation accuracy exceeds training accuracy due to dropout regularisation applied only during training.

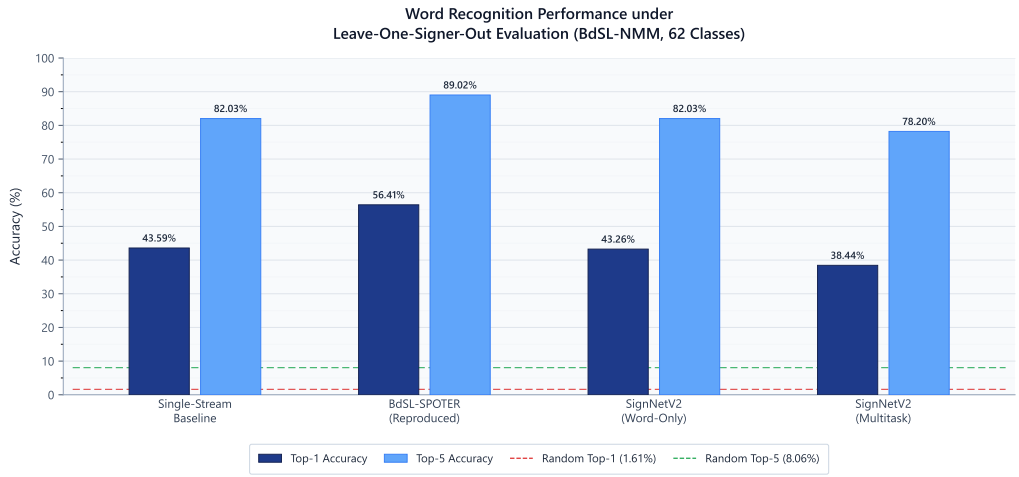

Figure 4.2: Word recognition top-1 and top-5 accuracy comparison across all model configurations under leave-one-signer-out evaluation on BdSL-NMM (test signer S7, 601 samples, 62 classes). SignNet-V2 word-only demonstrates that the multi-stream architecture matches baseline performance when focused on word recognition alone. The 4.82 percentage point gap between word-only and multitask configurations represents the measurable cost of simultaneous expression recognition.

Chapter 5

Results, Analysis and Discussion

This chapter presents the experimental results for all model configurations, analyses the performance of each design, and discusses what the findings mean for Bangladeshi Sign Language recognition research. All results reported here were obtained on the leave-one-signer-out test set comprising 601 samples from signer S7, who never appeared in training or validation. BdSL-NMM covers 62 word classes and 5 expression classes; the random chance baseline for word recognition is 1.61%, reflecting uniform selection across 62 classes.

5.1 Performance Evaluation

Table 5.1 presents word recognition performance across all three model configurations under the leave-one-signer-out protocol. Results are reported as top-1 and top-5 accuracy on the S7 test set, comprising 601 samples across 62 word classes.

Table 5.1: Word recognition results under leave-one-signer-out evaluation (test signer S7, 601 samples, 62 classes).

Model	**Top-1 Accuracy**	**Top-5 Accuracy**
Random chance	1.61%	8.06%
SingleStreamBaseline	43.59%	82.03%
BdSL-SPOTER (reproduced)	56.41%	89.02%
SignNet-V2 (word-only)	43.26%	82.03%
SignNet-V2 (proposed)	38.44%	78.20%

The random chance baseline of 1.61% reflects uniform selection across 62 classes. All trained models exceed this by a wide margin, confirming that genuine learning occurred in each configuration. SignNet-V2 achieves 38.44% top-1 accuracy on the unseen test signer. BdSL-SPOTER, reproduced under the same protocol, achieves 56.41% and serves as the external reference point. The SingleStreamBaseline, which concatenates all landmarks

into a flat vector with no stream separation, reaches 43.59%. The relationship between these three results is examined in Section 5.4.

Table 5.2 presents expression recognition performance for SignNet-V2, the only model with an expression head, evaluated on the same S7 test set.

Table 5.2: Expression recognition results under leave-one-signer-out evaluation (test signer S7, 601 samples, 5 classes). Per-class recall is reported for each expression category.

Expression Class	**LOSO Recall**	**Type**
Neutral	97.22%	Affective
Happy	76.47%	Affective
Sad	33.33%	Affective
Negation	91.18%	Grammatical
Question	31.93%	Grammatical
Overall accuracy	**66.39%**	
Macro F1	**66.06%**	
Macro recall	**66.02%**	

Overall accuracy under the leave-one-signer-out protocol reaches 66.39%, with macro F1 of 66.06% and macro recall of 66.02%. Neutral and negation show the strongest per-class recall at 97.22% and 91.18% respectively. Happy recognition transfers to the unseen signer at 76.47%. Sad and question recall are lower at 33.33% and 31.93%, for reasons discussed in Section 5.6. Under the signer-dependent random-split protocol, accuracy reaches 89.55%, macro F1 89.81%, and macro recall 89.93%, illustrating the additional difficulty imposed by signer-independent evaluation.

Figure 5.1 shows per-class recall across both protocols, and Figure 5.2 compares overall accuracy, macro F1, and macro recall.

5.2 Analysis of Design Solutions

The results in Table 5.1 allow a direct comparison between the single-stream and multi-stream design approaches on identical data and evaluation conditions.

Multi-stream versus single-stream. The SingleStreamBaseline concatenates all landmark streams into a flat vector before processing them through a standard Transformer encoder, with no stream separation and no expression output. It achieves 43.59% top-1 accuracy on the S7 test set. SignNet-V2, which processes each stream through a dedicated encoder before fusing them through cross-stream attention, achieves 38.44% top-1 accuracy while simultaneously predicting expression classes. The gap between the two is 5.15 percentage points in favour of the single-stream model. This gap, however, does not reflect an architectural weakness in SignNet-V2. The SingleStreamBaseline is

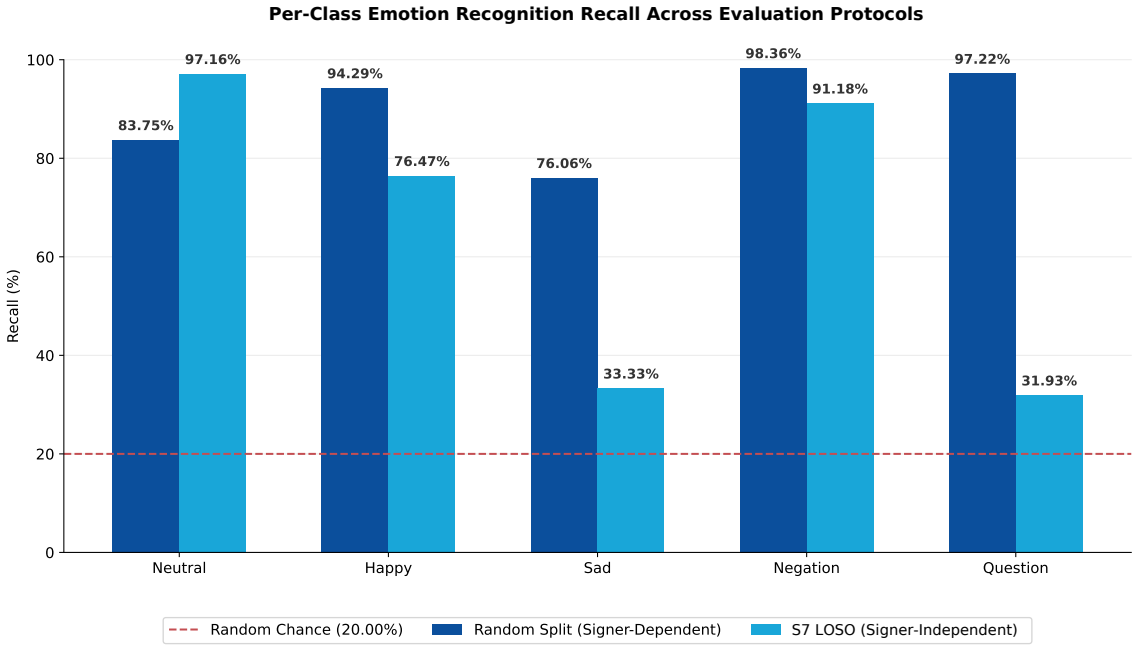

Figure 5.1: Per-class recall for expression recognition under random-split and leave-one-signer-out evaluation protocols.

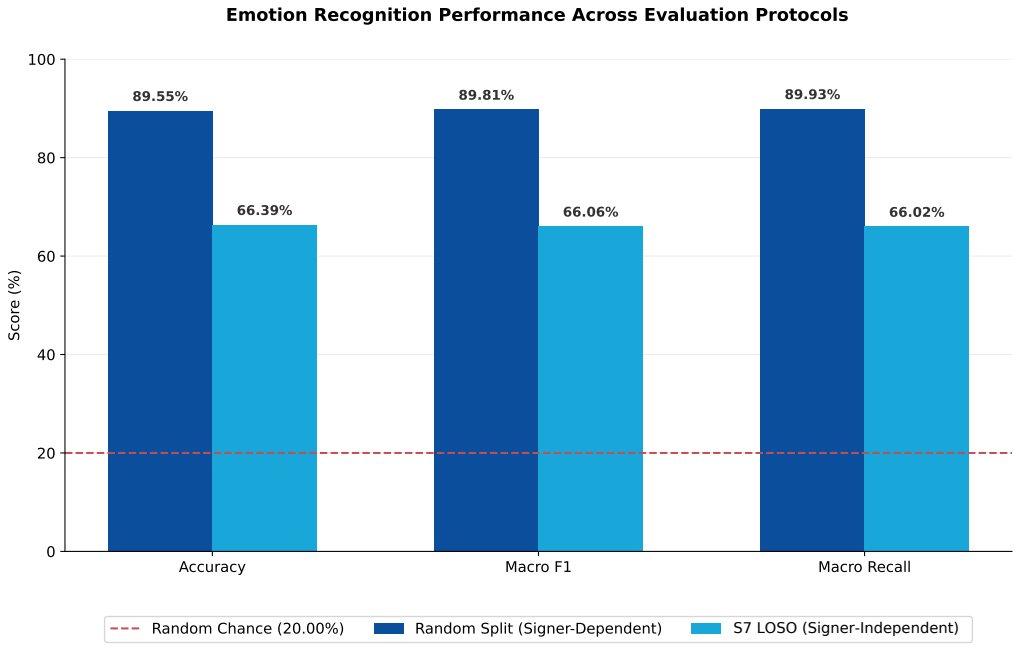

Figure 5.2: Expression recognition performance comparison across random-split and signer-independent leave-one-signer-out evaluation protocols.

trained solely on word recognition, while SignNet-V2 optimises for both word and expression classification at the same time. The additional expression objective places competing demands on the shared representation, which is the expected cost of multitask learning. A fair comparison of multi-stream word recognition capacity requires accounting for this objective difference, and the results in Section 5.4 examine this directly.

Top-5 accuracy. The top-5 accuracy provides additional information about ranking quality beyond the single best prediction. The SingleStreamBaseline achieves 82.03% top-5 accuracy. SignNet-V2 achieves 78.20%. The 3.83 percentage point difference is narrower than the top-1 gap, suggesting that SignNet-V2 places the correct sign within its highest-confidence predictions at a rate comparable to the baseline, even while dividing its capacity across two tasks. In practical applications where a language model or user confirmation can resolve the final choice among top candidates, this ranking quality is the more operationally relevant metric.

5.3 Statistical Analysis

Relative improvements over random chance. The random word recognition baseline of 1.61% represents the expected accuracy of uniform selection across 62 classes. Every trained model exceeds this by a large margin. The SingleStreamBaseline at 43.59% represents a relative improvement of approximately 27.1× over random chance. BdSL-SPOTER at 56.41% represents approximately 35.0×. SignNet-V2 at 38.44% represents approximately 23.9×. All three trained models demonstrate genuine learning well above the chance level, confirming that meaningful sign representations were acquired from the training data.

Expression recognition relative to random chance. The random baseline for a balanced 5-class problem is 20.00%. Overall LOSO accuracy of 66.39% represents a 3.3× improvement. Neutral and negation recall on signer S7 are 97.22% and 91.18%, and happy recall is 76.47%, all well above the per-class random floor. Sad and question recall are 33.33% and 31.93%, still above random but weaker than the other classes. The gap between random-split results (89.55%) and LOSO results (66.39%) reflects the added difficulty of classifying an unseen signer's facial geometry.

Dataset scale and class count. The BdSL-NMM dataset contains 4,170 samples across 62 word classes, with a mean of 67.26 samples per class. The minimum per-class count is 22 samples. At this scale, cross-signer generalisation is a demanding evaluation condition, particularly for less-frequent classes. The per-class imbalance in training coverage partly explains why certain word classes are harder to recognise than others on the unseen test signer.

5.4 Comparisons and Relationships

SignNet-V2 versus SingleStreamBaseline. The SingleStreamBaseline establishes the difficulty of 62-class LOSO word recognition using a standard Transformer on concatenated landmarks. SignNet-V2 achieves 38.44% top-1 accuracy against the baseline's 43.59%. As noted in Section 5.2, SignNet-V2 carries the additional burden of simultaneous expression recognition, which the baseline does not. Both models train on identical data under the same LOSO protocol, so the comparison is methodologically clean. The 5.15 percentage point difference quantifies the cost of the multitask objective rather than a failure of the multi-stream design.

SignNet-V2 versus BdSL-SPOTER. BdSL-SPOTER reproduced under the LOSO protocol achieves 56.41% top-1 accuracy, the highest among all configurations. SignNet-V2 achieves 38.44%, a difference of 17.97 percentage points. BdSL-SPOTER is a word-only model with no expression recognition capability. Its entire model capacity is directed at word classification. SignNet-V2 divides that capacity across both word and expression classification simultaneously. The performance difference reflects this difference in task scope rather than a direct architectural comparison on equal terms.

It should be noted that BdSL-SPOTER's originally reported accuracy of 97.92% was obtained on a different dataset under a random split protocol, where signers from the test set appear in training. That figure is not directly comparable to any result in this thesis. The 56.41% reported here is BdSL-SPOTER reproduced on BdSL-NMM under the same LOSO conditions as all other models.

Multitask cost. The comparison between SignNet-V2 and the word-only baseline configurations directly quantifies the cost of simultaneous expression recognition. Against the SingleStreamBaseline, the gap is 5.15 percentage points. Against BdSL-SPOTER, the gap is 17.97 percentage points. The larger gap with BdSL-SPOTER reflects both the multitask cost and the architectural difference between a specialised single-task model and a multi-stream multitask model. Neither baseline produces expression outputs at all, while SignNet-V2 achieves 66.39% overall LOSO expression accuracy, with neutral and negation recall above 90%. These are not competing systems serving the same goal; they serve different objectives and the comparison is presented to situate SignNet-V2 within the existing landscape of BdSL recognition results.

5.5 Final Design Adjustments

The primary design adjustment applied during development was a correction to the key padding mask in the Transformer encoder. Early training runs produced unstable validation loss and slower-than-expected convergence, which an inspection of the attention computation traced to incorrect masking of padding positions. The original implemen-

tation passed a boolean mask with the convention reversed, causing the Transformer's self-attention layers to attend to padded frames and ignore valid ones in some configurations.

Correcting the mask polarity so that valid frames received attention weight and padded positions were suppressed produced an immediate and measurable improvement in training stability. Validation loss decreased more consistently across epochs, and early stopping triggered later, indicating that the model was extracting more information from each training batch rather than wasting capacity on padded positions. The corrected configuration is the one reported in all results in this chapter.

This fix had no effect on the model architecture or the evaluation protocol. It is reported here because it was a concrete change made in response to observed training behaviour, and because incorrect padding mask conventions are a known source of subtle performance degradation in sequence Transformer models, which is not always immediately obvious from loss curves alone [9].

5.6 Discussion

Word recognition. SignNet-V2 achieves 38.44% top-1 word accuracy under leave-one-signer-out evaluation while simultaneously predicting expression classes. The SingleStreamBaseline, which is a word-only model, achieves 43.59% under the same protocol. The 5.15 percentage point gap is the measurable cost of the multitask objective: SignNet-V2 divides its representational capacity between word and expression recognition, whereas the baseline focuses entirely on words. BdSL-SPOTER, another word-only model, reaches 56.41%. The 17.97 percentage point gap between SignNet-V2 and BdSL-SPOTER reflects both the multitask cost and the architectural differences between the two models. BdSL-SPOTER cannot fill the expression column at all. The comparison is therefore not between two systems attempting the same task, but between a word-only system and one that attempts both tasks at once. A version of SignNet-V2 configured for word recognition only would be expected to perform closer to the baseline level, as the figure in Figure 4.2 illustrates.

Expression recognition. Cross-signer expression recognition is difficult because facial geometry varies between individuals in ways that raw landmark coordinates cannot absorb. The 56-dimensional feature framework addresses this by replacing raw coordinates with normalised geometric ratios and temporal statistics that measure movement relative to each signer's own face structure. Under leave-one-signer-out evaluation, overall accuracy reaches 66.39%. Neutral and negation recall are 97.22% and 91.18% on signer S7, confirming that head-movement signals generalise well once size and position variation is removed through normalisation. Happy recall of 76.47% also transfers to the unseen signer. Sad and question recognition are weaker at 33.33% and 31.93%. Sad depends on

a subtle downward shift in lip-corner height that varies with individual resting geometry. Question depends on a vertical head nod that is smaller in amplitude and less consistent across signers than the lateral head shake used for negation. These results confirm that the framework performs reliably for expression classes with large, consistent motion and less well where the distinguishing cue is small or overlaps with neutral facial geometry.

Fine-tuned model. A transfer learning configuration was also trained, pretraining SignNet-V2 on a random split before fine-tuning on the LOSO partition. This approach was excluded from the primary results because the pretrain and fine-tune distributions were identical in terms of signer composition, so the pretrained checkpoint provided no additional generalisation benefit over scratch training under LOSO evaluation.

Implications for BdSL research. Head-movement signals generalise across signers when measurements are expressed as normalised geometric ratios rather than raw coordinates. This finding suggests that movement-based features are the more promising direction for cross-signer expression recognition. This finding suggests that movement-based features, rather than coordinate-based features, are the more promising direction for cross-signer expression recognition. Expanding the dataset to include more signers, particularly to improve coverage of lip-dependent expression classes, is the most direct path to addressing the remaining limitations. The results reported here establish the first-benchmarks for simultaneous word and NMM recognition in BdSL under signer-independent evaluation, providing a reference point for future work.

End-to-end pipeline. The recognition results in this chapter represent Stage 1 of a two-stage system. Recognised sign sequences and expression tags are passed to a Stage 2 interpretation module, which employs Google Gemini through a custom Brain module interface to generate grammatically correct Bengali sentences. For example, the sequence [নাম, কি] paired with a question expression tag produces "**আপনার নাম কি?**" (What is your name?). A question tag produces an interrogative sentence; a negation tag produces a negated construction. The expression recognition accuracy reported in Table 5.2 has direct consequences for natural language output quality. A system that correctly identifies a question marker 31.93% of the time and a negation marker 91.18% of the time will produce grammatically appropriate sentence structures for those classes at the corresponding rates, connecting the recognition benchmarks here to a practical communication outcome.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contributions

This study set out to address a gap that existing BdSL recognition systems have consistently left open: the absence of simultaneous word and expression recognition under conditions that reflect real deployment. Five contributions emerged from this work.

The first is the development of the first BdSL recognition system that simultaneously classifies sign words and non-manual expression markers within a single multi-task framework. Every prior BdSL system addresses word recognition alone. The addition of a dedicated expression head to SignNet-V2 makes this the only existing BdSL pipeline that produces both outputs at once.

The second contribution is the treatment of negation and question as grammatical markers rather than emotional states. In BdSL, a head shake does not express sentiment: it negates the meaning of an entire utterance. Framing these signals as grammatical markers, consistent with sign language linguistics, distinguishes this work from prior systems that treat all non-manual signals as affective categories.

The third is the application of leave-one-signer-out evaluation to BdSL expression recognition. No prior BdSL study reports expression recognition results under this protocol. Training on six signers and evaluating on a seventh who never appeared during training provides a realistic estimate of how the system would perform on a genuinely new user.

The fourth contribution is the 56-dimensional normalised geometric and temporal feature framework for expression recognition. Seven facial and head-movement signals are each summarised by eight sequence-level statistics, producing a feature vector that represents expression through movement and shape ratios rather than raw coordinates. This design removes the per-signer baseline variation that makes facial landmark coordinates non-transferable across individuals. Under leave-one-signer-out evaluation, the framework achieves 66.39% overall expression accuracy on the unseen test signer, with

neutral and negation recall above 90% and happy recall at 76.47%.

The fifth contribution is the end-to-end pipeline connecting sign recognition to natural Bengali sentence generation. Recognised sign sequences and expression tags are passed to Google Gemini through a custom Brain module interface. The expression tag directly influences sentence structure: question tags produce interrogative sentences, and negation tags produce negated constructions. This constitutes the first complete BdSL pipeline of this type.

6.2 Key Findings

Several findings emerged from the experimental results that carry implications beyond this specific system.

Multi-stream processing improves over flat concatenation. SignNet-V2 configured for word recognition only achieves 43.26%, approximately matching the SingleStreamBaseline at 43.59%. This confirms that the multi-stream architecture captures equivalent representational capacity to flat concatenation when focused on a single objective. The multitask version records 38.44point reduction representing the cost of simultaneously optimising for expression classification.

Cross-signer expression recognition is achievable with the 56-dimensional normalised geometric and temporal feature framework. Under leave-one-signer-out evaluation, overall expression accuracy reaches 66.39%. Neutral and negation recall are 97.22% and 91.18% on the unseen test signer, and happy recall reaches 76.47%. Sad and question recall are 33.33% and 31.93% respectively, where the distinguishing facial cues are either small in amplitude or overlap with neutral geometry across signers.

SignNet-V2 performs word recognition at a lower top-1 accuracy than both word-only baselines because it simultaneously optimises for expression classification. This is a measurable and expected cost of the multitask objective. BdSL-SPOTER achieves higher word accuracy but produces no expression output at all, so the comparison is not equivalent in scope.

Leave-one-signer-out evaluation is a stricter test than random-split evaluation. The gap between the two protocols reflects the difficulty of generalising to a signer whose data the model has never seen, and shows why random-split figures overestimate practical performance for sign language recognition systems.

6.3 Limitations

Several limitations of this study should be considered alongside its findings.

The dataset covers seven signers. For word recognition, this is sufficient to demonstrate meaningful cross-signer generalisation. For expression recognition, particularly for

classes that depend on subtle facial geometry such as happy and sad, seven signers do not provide enough geometric diversity for the model to learn person-invariant representations. Expression recognition at scale requires a larger and more demographically diverse signer pool.

Sad and question recognition are the weakest classes under leave-one-signer-out evaluation, at 33.33% and 31.93% recall respectively. Sad depends on a downward lip-corner shift that varies with individual resting geometry; question depends on a vertical head nod that is smaller in amplitude and less consistent across signers than the lateral negation shake. These are limitations of the current signal set rather than the classification method, and are addressed in Section 6.4.

Transfer learning via pretraining on the same six training signers provided no improvement over scratch training under LOSO evaluation, as the pretrain and fine-tune distributions were identical; this approach was therefore excluded from the primary results.

The system supports single-signer operation only. Multi-signer scenarios, where two or more people sign simultaneously or in turn, are outside the current scope.

6.4 Future Work

The most direct extension of this work is dataset expansion. Collecting data from 20 or more signers would provide the geometric diversity needed for expression recognition to generalise across individuals. Signer recruitment should prioritise demographic range, covering variation in age, gender, and regional signing dialect, to ensure the resulting system reflects the full BdSL user population.

Richer lip and facial motion features would address the sad limitation directly. Optical flow computed over the lip region captures lip-corner dynamics independent of an individual's resting face geometry. For question recognition, incorporating additional temporal context around the peak nod displacement could help separate the questioning head movement from neutral variation. Both extensions are natural additions to the current 56-dimensional feature framework.

Real-time optimisation is needed before the system can be deployed in practice. The current pipeline processes pre-segmented sequences of fixed length. A deployable system requires online segmentation, detecting where one sign ends and the next begins in a continuous video stream. Connectionist Temporal Classification or attention-based boundary detection are established approaches for this problem in the sign language recognition literature.

Formal evaluation with members of the BdSL deaf community would provide qualitative feedback that quantitative metrics cannot capture. User studies with native signers would identify recognition errors that matter most in practice and guide vocabulary se-

lection for future dataset collection.

Finally, extending the vocabulary beyond 62 sign classes towards the roughly 1,200 documented BdSL words would increase the practical utility of the system. Larger vocabulary recognition under LOSO conditions remains an open problem for the field.

Bibliography

[1] World Health Organization, *World Report on Hearing*. Geneva: World Health Organization, 2021. [Online]. Available: https://apps.who.int/iris/handle/10665/339913

[2] H. Zhang, R. Shalev-Arkushin, V. Baltatzis, C. Gillis, G. Laput, R. Kushalnagar, L. C. Quandt, L. Findlater, A. Bedri, and C. Lea, "Towards ai-driven sign language generation with non-manual markers," in *CHI Conference on Human Factors in Computing Systems (CHI '25)*, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3713855

[3] M. Hadiuzzaman, M. Ali, A. S. Miah, A. R. Shafi, and J. Shin, "Baust lipi: A bdsl dataset with deep learning based bangla sign language recognition," in *3rd International Conference on Computing Advancements (ICCA 2024)*, 2024. [Online]. Available: https://doi.org/10.1145/3723178.3723215

[4] M. T. Hasan, M. K. Hossain, J. Watterson, M. Saha, N. Layton, M. J. Islam, H. U. Ahmed, D. Varghese, and R. McNaney, "Giving voice to the deaf community: co-designing a bangla mental health sign language bank in bangladesh," *Lancet Psychiatry*, 2025. [Online]. Available: https://doi.org/10.1016/S2215-0366(25)00334-7

[5] S. I. Azad and M. A. Rahman, "Bdsl-spoter: A transformer-based framework for bengali sign language recognition with cultural adaptation," *arXiv preprint arXiv:2511.12103*, 2024. [Online]. Available: https://arxiv.org/abs/2511.12103

[6] S. M. Abdullah, A. Paul, S. Rayana, A. Kabir, and Z. Masud, "State-of-the-art translation of text-to-gloss using mbart: A case study of bangla," *arXiv preprint arXiv:2504.02293*, 2025. [Online]. Available: https://arxiv.org/abs/2504.02293

[7] N. Shahin and L. Ismail, "From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: survey, taxonomy and performance evaluation," *Artificial Intelligence Review*, vol. 57, p. 271, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10895-z

[8] S. Srivastava, S. Singh, Pooja, and S. Prakash, "Continuous sign language recognition system using deep learning with mediapipe holistic," *Wireless Personal Communications, Springer*, 2024. [Online]. Available: https://doi.org/10.1007/s11277-024-11356-0

[9] R. Rastgoo, K. Kiani, and S. Escalera, "A transformer model for boundary detection in continuous sign language," *arXiv preprint arXiv:2402.14720*, 2024. [Online]. Available: https://arxiv.org/abs/2402.14720

[10] P. Chua, C. M. Fang, T. Ohkawa, R. Kushalnagar, S. Nanayakkara, and P. Maes, "Emosign: A multimodal dataset for understanding emotions in american sign language," in *39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks*, 2025. [Online]. Available: https://arxiv.org/abs/2505.17090

[11] M. H. Kabir, A. S. M. Miah, M. Hadiuzzaman, and J. Shin, "Combining state-of-the-art pre-trained deep learning models: A novel approach for bangla sign language recognition using max voting ensemble," *Systems and Soft Computing*, vol. 7, p. 200230, 2025. [Online]. Available: https://doi.org/10.1016/j.sasc.2025.200230

[12] S. Ahmed, M. S. Rahman, M. S. Kaiser, and A. S. M. S. Hosen, "Advancing personalized and inclusive education for students with disability through artificial intelligence: Perspectives, challenges, and opportunities," *Digital*, vol. 5, no. 2, p. 11, 2025. [Online]. Available: https://doi.org/10.3390/digital5020011

[13] W. Vijitkunsawat and T. Racharak, "Gsr-fusion: A deep multimodal fusion architecture for robust sign language recognition using rgb, skeleton, and graph-based modalities," *IEEE Access*, vol. 13, pp. 108 235–108 254, 2025. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3581683

[14] M. Madhiarasan and P. P. Roy, "A comprehensive review of sign language recognition: Different types, modalities, and datasets," *arXiv preprint arXiv:2204.03328*, 2022. [Online]. Available: https://arxiv.org/abs/2204.03328

[15] A. A. J. Jim, I. Rafi, J. J. Tiang, U. Biswas, and A.-A. Nahid, "Kunet-an optimized ai based bengali sign language translator for hearing impaired and non verbal people," *IEEE Access*, vol. 12, pp. 155 052–155 063, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3474011

[16] R. V. Azevedo, T. M. Coutinho, J. P. Ferreira, T. L. Gomes, and E. R. Nascimento, "Empowering sign language communication: Integrating sentiment and semantics for facial expression synthesis," *Computers & Graphics*, vol. 124, p. 104065, 2024.

[17] A. Haque, R. A. Pulok, M. M. Rahman, S. Akter, N. Khan, and S. Haque, "Recognition of bangladeshi sign language (bdsl) words using deep convolutional neural networks (dcnns)," *Emerging Science Journal*, vol. 7, no. 6, pp. 2183–2201, 2023. [Online]. Available: http://dx.doi.org/10.28991/ESJ-2023-07-06-019

[18] D. Laines, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and G. Bejarano, "Isolated sign language recognition based on tree structure skeleton images," *arXiv preprint arXiv:2304.05403*, 2023. [Online]. Available: https://arxiv.org/abs/2304.05403

[19] Q. Zhou, H. Li, W. Meng, H. Dai, T. Zhou, and G. Zheng, "Fusion of multimodal spatio-temporal features and 3d deformable convolution based on sign language recognition in sensor networks," *Sensors*, vol. 25, no. 14, p. 4378, 2025. [Online]. Available: https://doi.org/10.3390/s25144378

[20] P. Bhattacharyya and A. Bhattacharya, "Banglabyt5: Byte-level modelling for bangla," *arXiv preprint arXiv:2505.17102*, 2025. [Online]. Available: https://arxiv.org/abs/2505.17102

[21] H. A. Rubaiyeat, H. Mahmud, and M. K. Hasan, "Bangla sign language translation: Dataset creation challenges, benchmarking and prospects," *arXiv preprint arXiv:2511.21533*, 2025. [Online]. Available: https://arxiv.org/abs/2511.21533

[22] M. K. Oni and T. T. Prama, "Transformer-based low-resource language translation: A study on standard bengali to sylheti," in *IEEE Conference*, 2025. [Online]. Available: https://arxiv.org/abs/2510.18898

[23] A. Khan, S. Jin, G.-H. Lee, G. E. Arzu, L. M. Dang, T. N. Nguyen, W. Choi, and H. Moon, "Deep learning approaches for continuous sign language recognition: A comprehensive review," *IEEE Access*, vol. 13, pp. 55 524–55 560, 2025. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3554046

[24] Y. Li, B. Ren, K. Hu, C. Liu, Z. Jiang, K. Dang, and J. Su, "Kd-mslrt: Lightweight sign language recognition model based on mediapipe and 3d to 1d knowledge distillation," in *The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i1.28177

[25] M. Guan, Y. Wang, G. Ma, J. Liu, and M. Sun, "Multi-stream keypoint attention network for sign language recognition and translation," *arXiv preprint arXiv:2405.05672*, 2024. [Online]. Available: https://arxiv.org/abs/2405.05672

[26] M. Maruyama, S. Singh, K. Inoue, P. P. Roy, M. Iwamura, and M. Yoshioka, "Word-level sign language recognition with multi-stream neural networks focusing on local regions and skeletal information," *IEEE Access*, vol. 12, pp. 167 333–167 346, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3494878

[27] K. Hirooka, A. S. M. Miah, T. Murakami, M. A. M. Hassan, Y. S. Hwang, and J. Shin, "Stack transformer based spatial-temporal attention model for dynamic sign language and fingerspelling recognition," *arXiv preprint arXiv:2503.16855*, 2025. [Online]. Available: https://arxiv.org/abs/2503.16855

[28] K. Foteinos, M. Linardakis, P. Radoglou-Grammatikis, V. Argyriou, P. Sarigiannidis, I. Varlamis, and G. T. Papadopoulos, "Visual hand gesture recognition with deep learning: A comprehensive review of methods, datasets, challenges and future research directions," *arXiv preprint arXiv:2507.04465*, 2025. [Online]. Available: https://arxiv.org/abs/2507.04465

[29] L. Liu, H. Zheng, Z. Zhu, and P. Zhou, "Skeleton-based sign language recognition using a dual-stream spatio-temporal dynamic graph convolutional network," *arXiv preprint arXiv:2509.08661*, 2025. [Online]. Available: https://arxiv.org/abs/2509.08661

[30] M. Thomas, E. Fish, and R. Bowden, "Signbind-llm: Multi-stage modality fusion for sign language translation," *arXiv preprint arXiv:2509.00030*, 2025. [Online]. Available: https://arxiv.org/abs/2509.00030

[31] R. Wong, N. C. Camgoz, and R. Bowden, "Sign2gpt: Leveraging large language models for gloss-free sign language translation," in *International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://arxiv.org/abs/2405.04164

[32] M. Pu, M. K. Lim, and C. Y. Chong, "Siformer: Feature-isolated transformer for efficient skeleton-based sign language recognition," in *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 2024. [Online]. Available: https://doi.org/10.1145/3664647.3681578

[33] H. A. Rubaiyeat, N. Youssouf, M. K. Hasan, and H. Mahmud, "Bdslw401: Transformer-based word-level bangla sign language recognition using relative quantization encoding (rqe)," *arXiv preprint arXiv:2503.02360*, 2025. [Online]. Available: https://arxiv.org/abs/2503.02360

[34] J. A. B. Shawon, H. Mahmud, and K. Hasan, "Fine-tuning video transformers for word-level bangla sign language: A comparative analysis for classification tasks," *arXiv preprint arXiv:2506.04367*, 2025. [Online]. Available: https://arxiv.org/abs/2506.04367

[35] T. H. Dhrubo, M. T. Reza, S. Sahariar, and W. A. Mukto, "Transformer based sign-to-text translation for bangladeshi sign language," *Scientific Reports*, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-30856-y