DEEP LEARNING-BASED ARABIC SIGN LANGUAGE RECOGNITION FOR MEDICAL EMERGENCY COMMUNICATION

supervised by: Dr. Noor Felemban

Group 6

MOTIVATION & PROBLEM

- Deaf and hard-of-hearing individuals face major communication challenges during medical emergencies.
- Miscommunication or delays in emergencies can lead to serious risks.
- Arabic Sign Language recognition is still limited, especially compared to other sign languages.
- Most existing systems do not focus on emergency-related signs or real-world scenarios.
- Many models struggle to generalize to new signers and are not suitable for real-time use.

RITAJ

PROJECT GOALS

Main goal

to develop an effective deep learning model for Arabic Sign Language recognition in the context of medical emergencies.

Specific Goals

- Select a set of emergency-related Arabic signs that are relevant to critical situations.
- Compare different deep learning models for sign language recognition.
- Evaluate models under realistic settings, including signer-independent scenarios.
- Identify a model that balances accuracy, efficiency, and practicality for real-world deployment.

DATASET & SUBSET

- **Dataset:** KArSL-502 (word-level ArSL)
- **Full dataset:** 502 signs, 3 signers, 50 repetitions each → 75,300 samples
- **Selected subset:** 131 emergency-related signs (medically relevant vocabulary)
- **Subset size:** 131 × 3 × 50 = 19,650 samples
- Frames are stored as RGB sequences per repetition (video as frame folders)

snippet of the dataset frames

NORA

ORGANIZATION & LABEL MAPPING

- **Dataset structure:** signer folders 01/02/03, each with train/test splits
- **Custom Excel mapping created for the selected subset:**
  - ClassIndex (0–130)
  - SignFolder (zero-padded folder ID)
  - SignID, Arabic meaning, English translation
- Verified paths, repetitions, and frame readability; visually inspected samples

ClassIndex	SignFolder	SignID	Sign-Arab	Sign-English
0	0071	71	هيكل عظمي	Skeleton
1	0072	72	جمجمة	skull
2	0073	73	عمود فقري	Backbone
3	0074	74	قفص صدري	Chest
4	0075	75	جهاز تنفسي	Respiratory device
5	0076	76	قصبة هوائية	Trachea
6	0077	77	رئتان	lungs
7	0078	78	شهيق - زفير	Ins and Outs
8	0079	79	جهاز هضمي	digestive system
9	0080	80	وجه	Face
10	0081	81	بلعوم	pharynx
11	0082	82	كبد	liver
12	0083	83	البنكرياس	pancreas

snippet of Excel selected words

NORA

EXPERIMENTS & EVALUATED MODELS

- **ResNet + BiLSTM:** CNN for spatial feature extraction followed by BiLSTM for temporal modeling.
- **I3D (Inflated 3D Convolutional Network):** 3D convolutional network that learns spatiotemporal features directly from video.
- **VideoMAE:** Transformer-based model applied to RGB video frames for temporal representation learning.
- **MediaPipe + SignBart (Proposed):** Skeleton-based pipeline using extracted keypoints instead of video frames.

MOZOON

EVALUATION METRICS & RESULTS

Evaluation Metrics: Accuracy, Precision, Recall, and Macro F1-Score.

All models were evaluated using the same data split and test set.

- **VideoMAE:** Highest performance (~96–97% across metrics)
- **I3D:** Second-best results (~85–88%)
- **ResNet + BiLSTM:** Strong precision and recall (~82-87%)
- **MediaPipe + SignBart:** ~80% across metrics under more challenging conditions

MOZOON

WHY MEDIAPIPE?

We focused on medical emergency communication, where systems must be fast, lightweight, privacy-aware, and generalizable.

MediaPipe extracts body and hand skeletal keypoints instead of raw video:

- This makes the system:
  - Computationally efficient
  - Privacy-preserving (no raw images stored)
  - Suitable for real-time deployment
- It reduces input size (dimensionality) while preserving essential motion information

SHADEN

SIGNBART

- SignBart is a transformer-based encoder–decoder model
- It is designed specifically for skeleton-based sign language recognition
- Unlike other video models, it:
  - Focuses on temporal motion patterns
  - Uses attention to model long-range dependencies across frames
  - This makes it well suited for sign understanding

SHADEN

PIPELINE

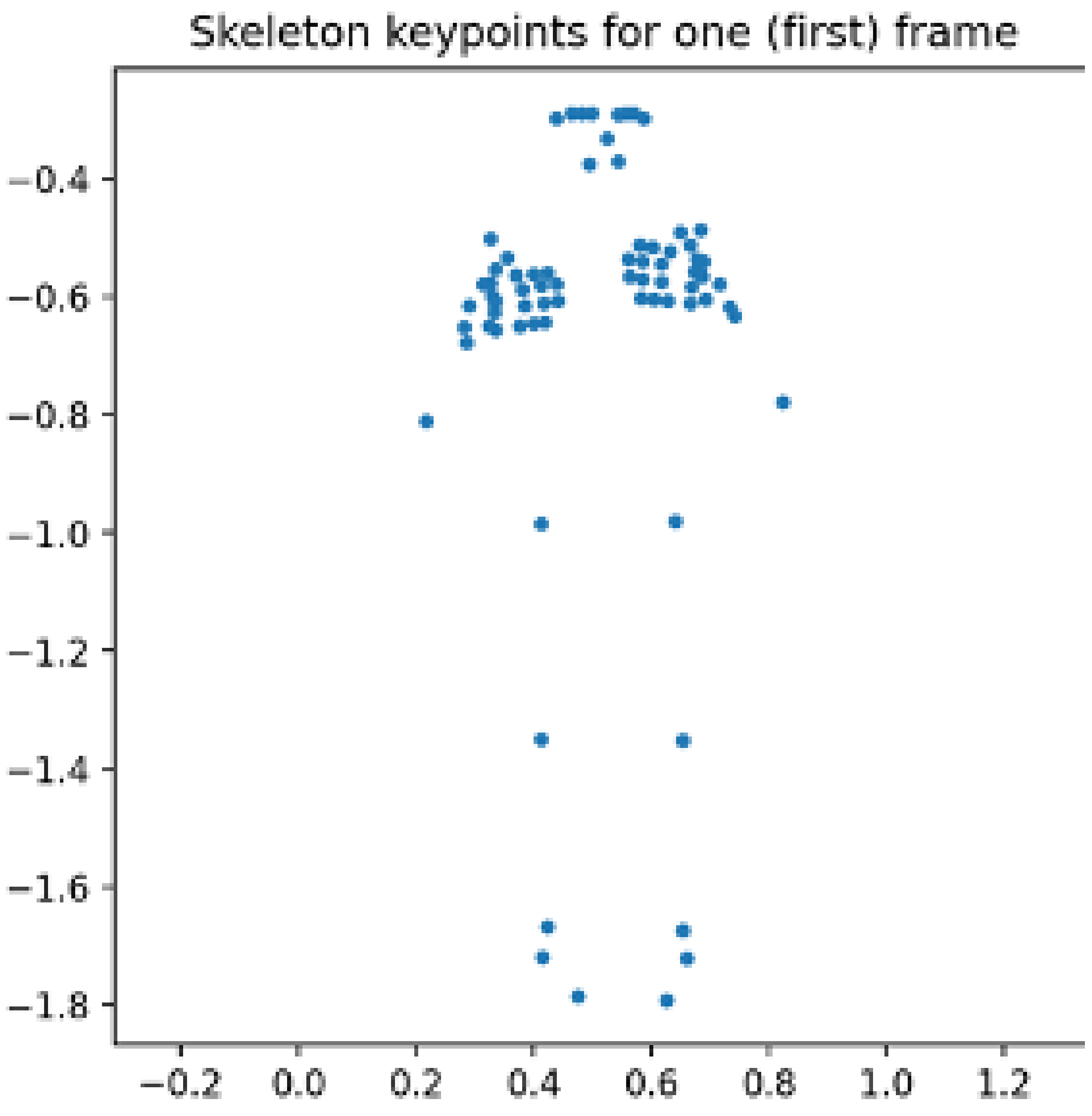
- Input sign videos
- Extract body + hand keypoints using MediaPipe
- Convert them into fixed-length sequences (50 frames)
- Feed them into SignBart
- Output one of 131 medical emergency sign classes

- Uses 75 keypoints per frame
- Relies only on motion and structure
- Ignores background, clothing, and signer identity (joints only)

SHADEN

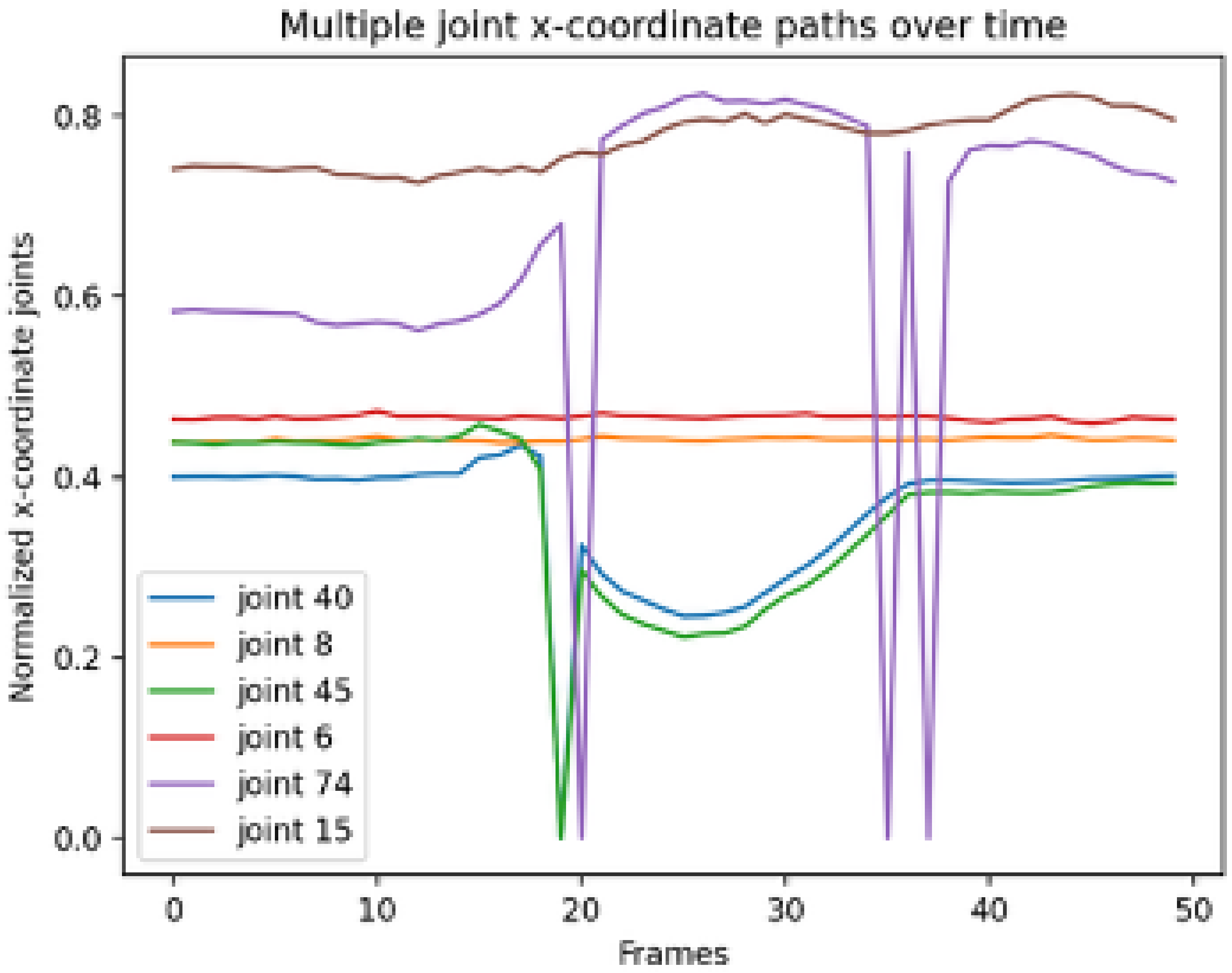
Framework	PyTorch
Optimizer	AdamW
Learning Rate	3×10-4
Weight Decay	1×10-2
Batch Size	64
Maximum Epochs	200
Learning Rate Scheduler	Cosine Annealing
Gradient Clipping	Max gradient norm = 1.0
Loss Functio	Cross-Entropy Loss
Validation Frequency	After each training epoch
Early Stopping	Patience of 15 epochs
Model Selection	Best validation accuracy checkpoint

SHADEN



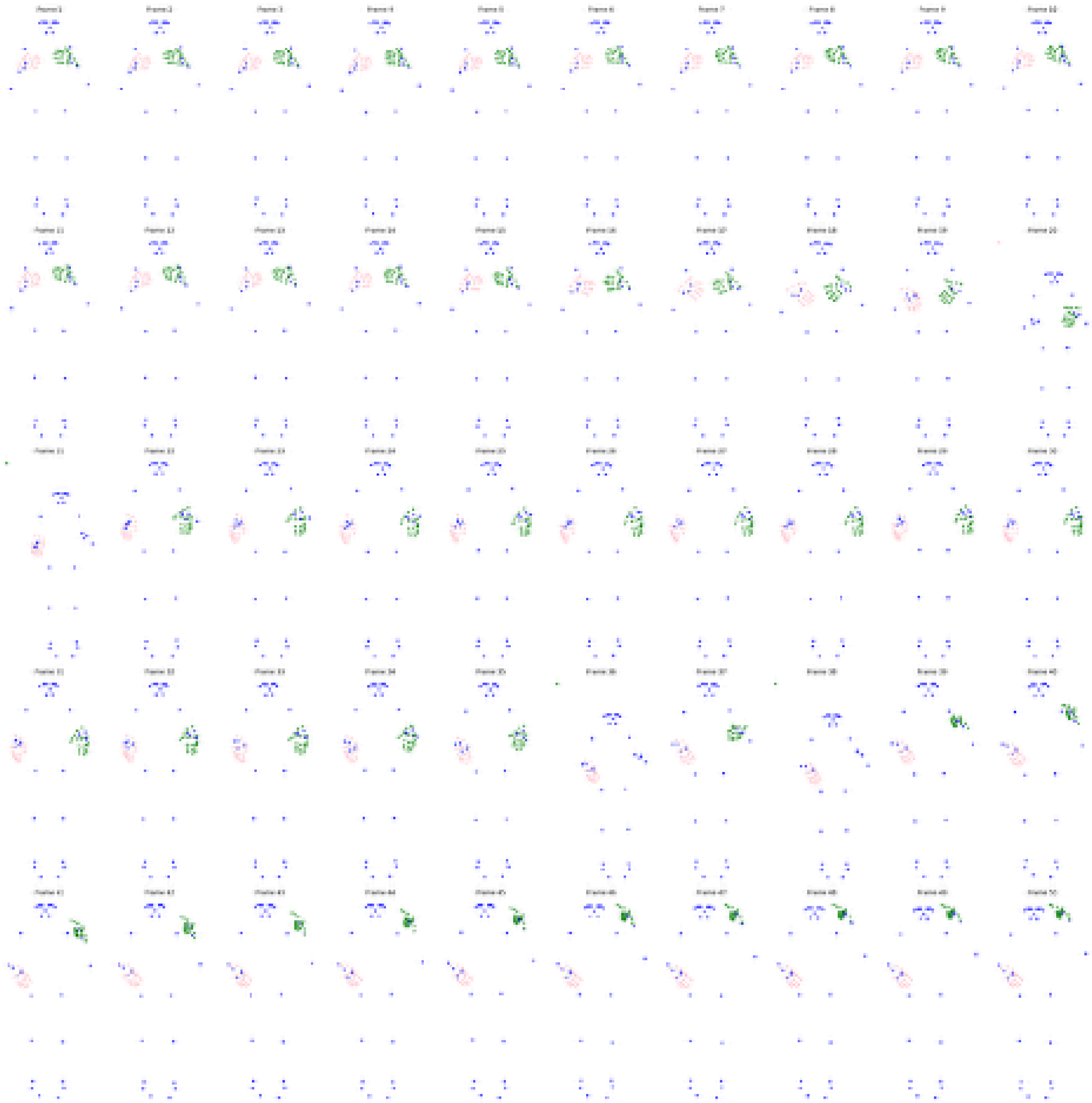
Example of extracted skeletal keypoints for a single frame showing body pose, left-hand, and right-hand landmarks.

SHADEN



Some joints -coordinates trajectories across consecutive frames

SHADEN



Skeleton visualization across 50 frames for a sample sign

SHADEN

LIMITATIONS

Limited Training Data

Dataset size was reduced for feasibility, impacting overall model performance.

Controlled Environment

Reliance on fixed lighting and professional signers limits real-world generalization

Restricted Input Modality

The skeleton-based model excludes RGB data, omitting visual details like hand shape

HANEEN

CONCLUSION & FUTURE WORK

- we evaluated five models and VideoMAE achieved the highest accuracy 91% in signer-independent testing.
- MediaPipe + SignBart proved optimal for real-time, resource-constrained applications.
- Results confirm that modeling temporal dynamics is essential for accurate ArSL recognition.
- Future work aims to improve generalization by expanding the dataset with more signs and signers.

HANEEN

THANK YOU!

Prepared By:
Shaden Alsuhayan
Haneen Alhabib
Mozoon Alkhalis
Haneen Alomri
Nora Aljomuh
Ritaj Alhamli