WQF7006 COMPUTER VISION AND IMAGE PROCESSING

GROUP ASSIGNMENT

MALAYSIAN SIGN LANGUAGE TRANSLATION CASE STUDY

GROUP INFORMATION

- OCC 1, GROUP 6
- GROUP MEMBERS:
  1. Chee Zen Yu (24088354)
  2. Cheong Yi Fong (U2005327)
  3. Hong Jia Herng (U2005313)
  4. Khor Yin Loon (23115881)

TEAM ROLES

Khor Yin Loon
Project Manager

Planned and coordinated the project workflow, consolidated the final presentation and report

Hong Jia Herng
AI Engineer

Preprocess dataset, trained deep learning model and evaluated its performance

Cheong Yi Fong
AI Engineer

Experiment with different models, and perform model comparison

Chee Zen Yu
Software Developer

Integrated the trained model into a simple application workflow

INTRODUCTION

- Malaysian Sign Language (MSL) is an important communication tool for the **deaf and hard-of-hearing** community in Malaysia [1]
- Automatic sign language recognition can improve accessibility and inclusive communication
- Developing AI models for MSL is challenging due to **limited datasets and low-resource language conditions**
- This project explores the use of computer vision and deep learning to recognize MSL glosses
- Our goal is to design and demonstrate a prototype MSL recognition system with real-world usability

PROBLEM STATEMENT

- **Communication between MSL users and non-signers** remain a major challenge in daily life
- Most existing sign language recognition systems focus on American Sign Language (ASL) or other high-resource languages [2]
- MSL **lacks large annotated datasets** and practical AI-based tools
- As a result, accessibility and inclusive communication in Malaysia remain limited

TARGET USERS & USE CASE EXAMPLES

- Deaf and hard-of-hearing MSL users:
  - » Use the app to translate their sign language into text for basic communication with non-MSL users in daily situations.
- Hearing individuals learning or interacting with MSL users
  - » Use the app to understand and learn MSL signs during interactions or practice sessions.
- Teachers and students in special education and MSL learning environments
  - » Use the app as a teaching and learning aid to demonstrate, recognize, and practice MSL glosses.

DATA COLLECTION

- Data collection was conducted collectively with all course participants during scheduled sessions
- Malaysian Sign Language videos were recorded under instructor guidance
- A shared dataset containing multiple MSL glosses was created across all teams
- Videos were recorded in a controlled environment to ensure consistent framing and visibility
- Our group **selected top 30 glosses with the most samples** to ensure **sufficient data per class** while **maintaining model robustness**

DATA COLLECTION

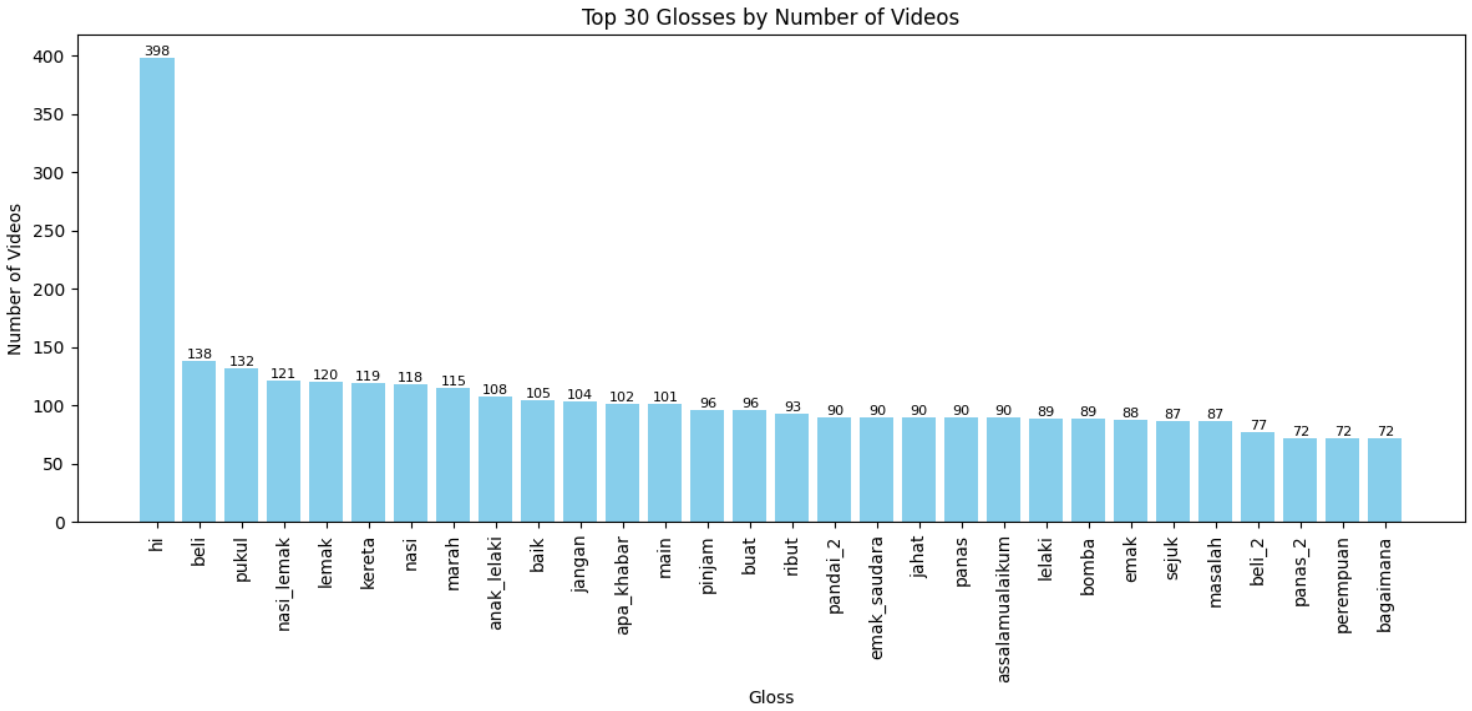


Figure shows top 30 selected glosses by number of videos:

- ['hi', 'beli', 'pukul', 'nasi_lemak', 'lemak', 'kereta', 'nasi', 'marah', 'anak_lelaki', 'baik', 'jangan', 'apa_khabar', 'main', 'pinjam', 'buat', 'ribut', 'pandai_2', 'emak_saudara', 'jahat', 'panas', 'assalamualaikum', 'lelaki', 'bomba', 'emak', 'sejuk', 'masalah', 'beli_2', 'panas_2', 'perempuan', 'bagaimana']

LANDMARK EXTRACTION & DATA PREPROCESSING

- Filtered sign language videos frame-by-frame with MediaPipe Hollistic model on detection of pose and hand landmarks
- **Uniformly sampled** 30 frames throughout filtered frames
- Combined extracted pose and hand landmarks into a unified feature vector
- Assigned each gesture with a numeric label and organized into sequences for model training

MODEL ARCHITECTURE

- Baseline LSTM
  - A stacked LSTM model that learns temporal patterns in hand and pose landmark sequences by modeling frame-to-frame dependencies.
- BiLSTM with Attention
  - An enhanced LSTM architecture that processes sequences in both forward and backward directions and applies attention to focus on the most informative frames.
- Transformer Encoder
  - A self-attention-based model that captures global temporal relationships across all frames without relying on recurrent connections.

Model	No. of Parameters
Baseline	491,806
BiLSTM+Attention	829,599
Transformer	301,982

BASELINE MODEL TRAINING

- A provided baseline deep learning model was used as the foundation, with a custom LSTM implemented to capture temporal gesture patterns
- The model was trained on the top 30 MSL glosses with the highest number of samples
- The dataset was split into 90% for training and 10% for testing
- Training with:
  - Early stopping
  - Gradient clipping

Hyperparameter	Label
Optimizer	Adam
Learning rate	1e-3
Epochs	100
Batch size	32
Loss	Categorical Cross-entropy

FINDINGS: Selecting the number of classes

Number of classes	Test Accuracy
10	0.3252
20	0.5328
30	0.473

- Initial experiments were conducted using 10, 20, and 30 gesture classes to study the effect of class size on model performance
- Among baseline models, the 20-class configuration achieved the highest initial accuracy, indicating a better balance between data availability and task complexity

FINDINGS: Selecting the number of classes

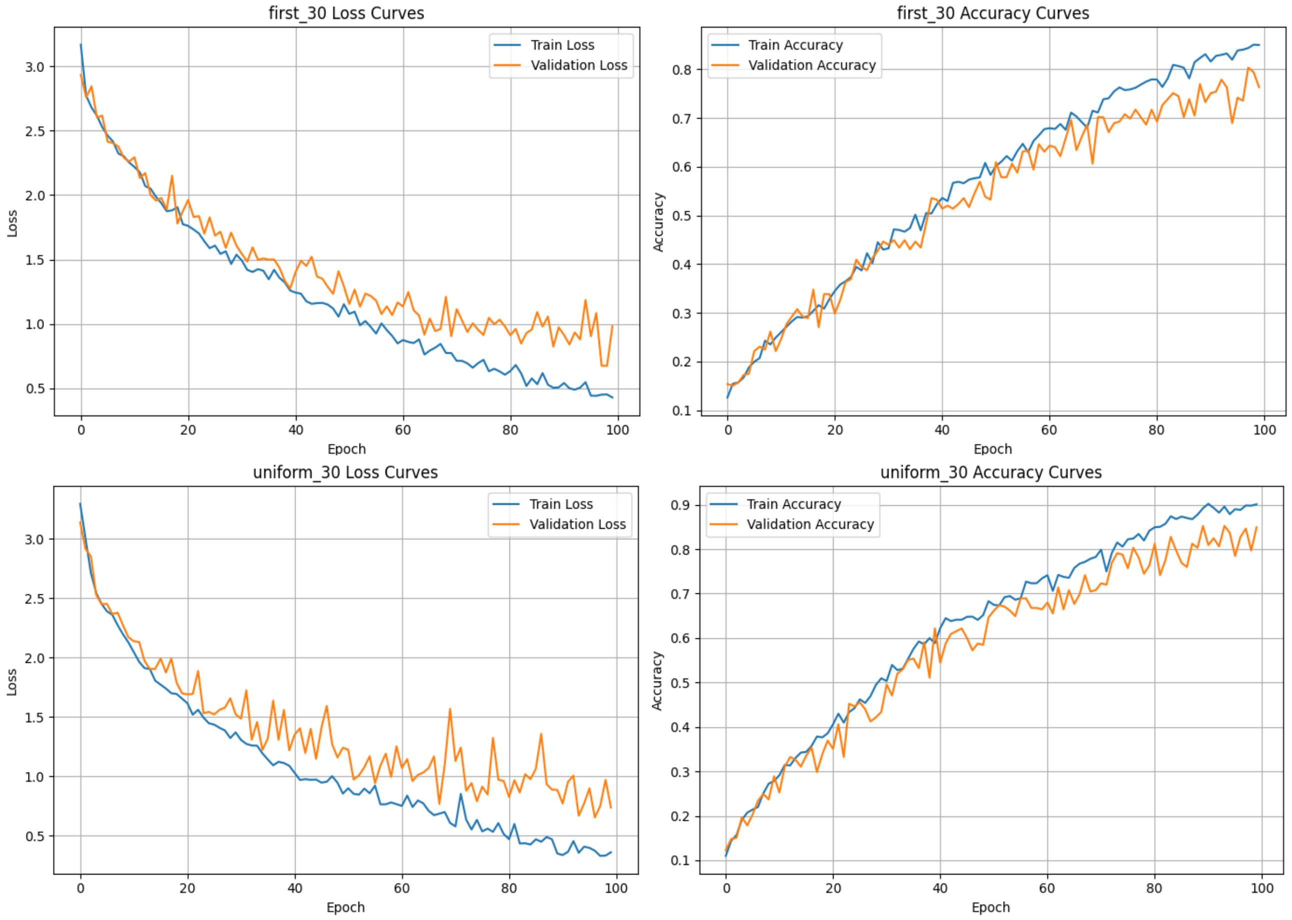
Number of classes	Test Accuracy
10	1
20	0.918
30	0.9509

What we did?

- Train longer
- Gradient clipping for stable learning

- The optimized 10-class model achieved very high accuracy but showed signs of **overfitting**
- The 30-class model with full optimization achieved the best overall performance, demonstrating strong generalization
- We hypothesize that training on a larger number of gesture classes increases feature diversity, which helps improve discrimination across gestures

FINDINGS: Which Preprocessing Strategy is better? First 30 Frames vs. Uniformly Sample 30 Frames



- We trained the baseline model with 2 different preprocessing strategies:

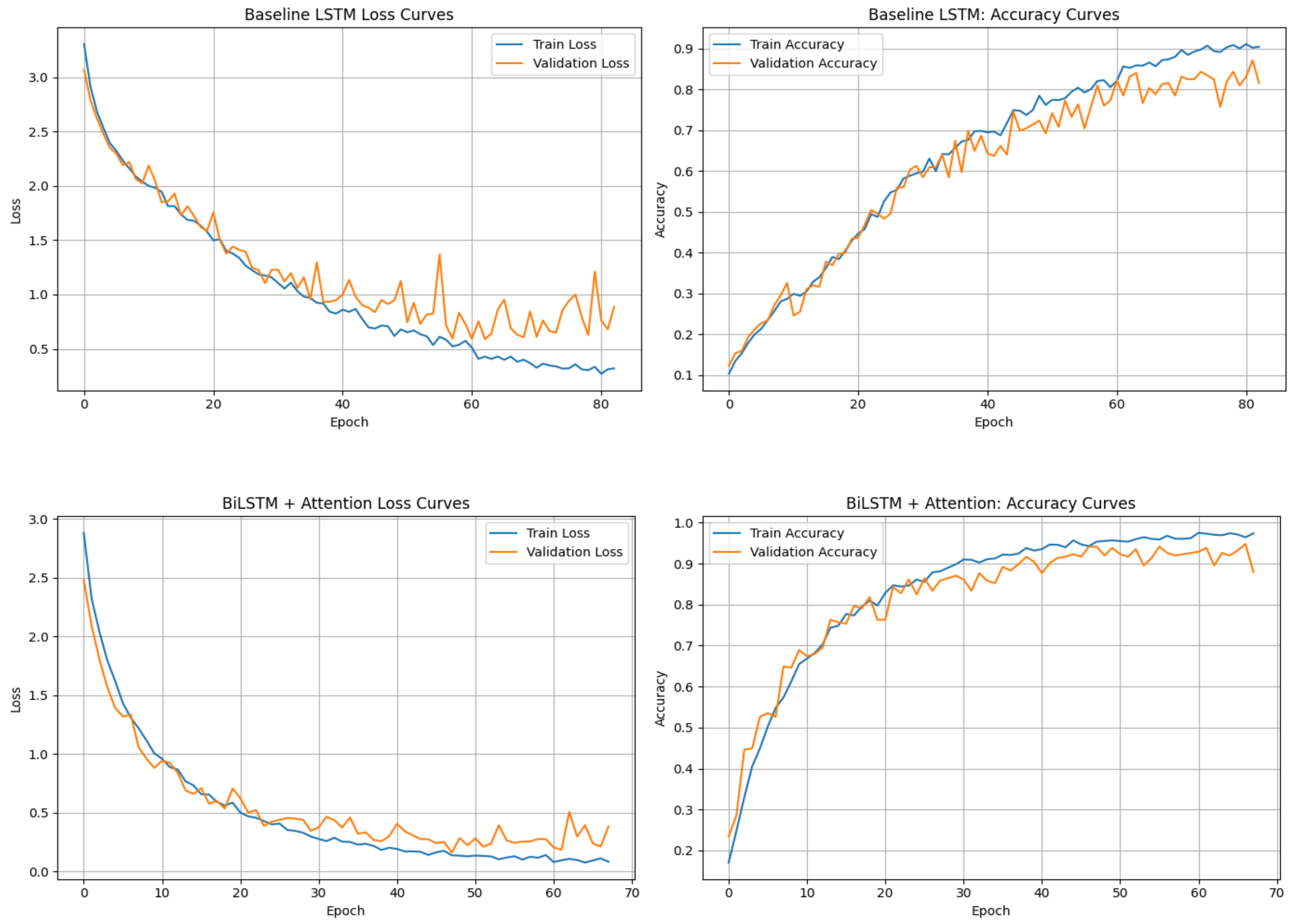
Strategy	Test Loss	Test Accuracy
First_30	0.9803	0.7631
Uniform_30	**0.7391**	**0.8492**

FINDINGS: Baseline LSTM vs. BiLSTM+Attention vs. Transformer

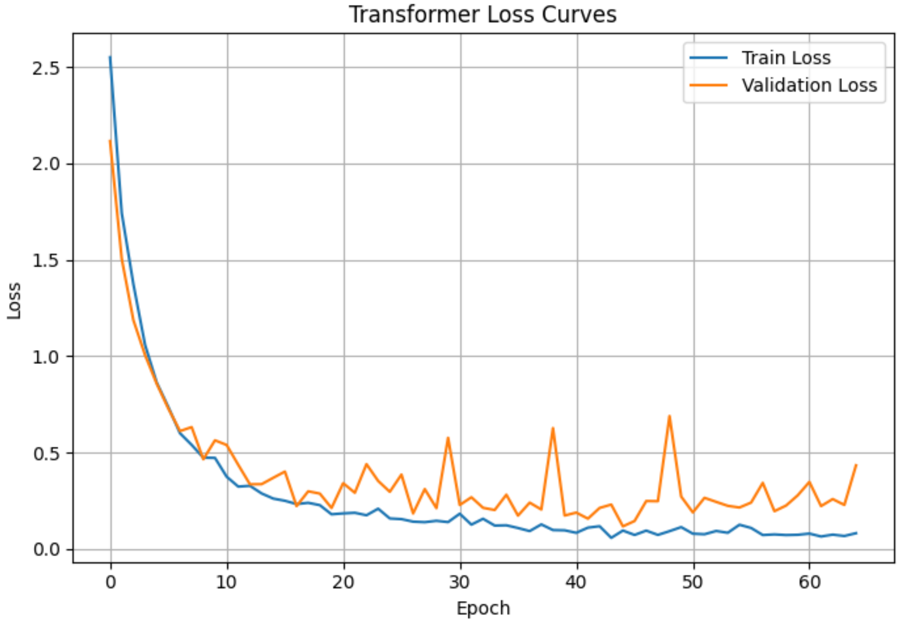
Model	Test Accuracy	Test F1 Score
Baseline LSTM	0.8308	0.8247
BiLSTM + Attention	0.9415	0.9413
Transformer	**0.9815**	**0.9814**

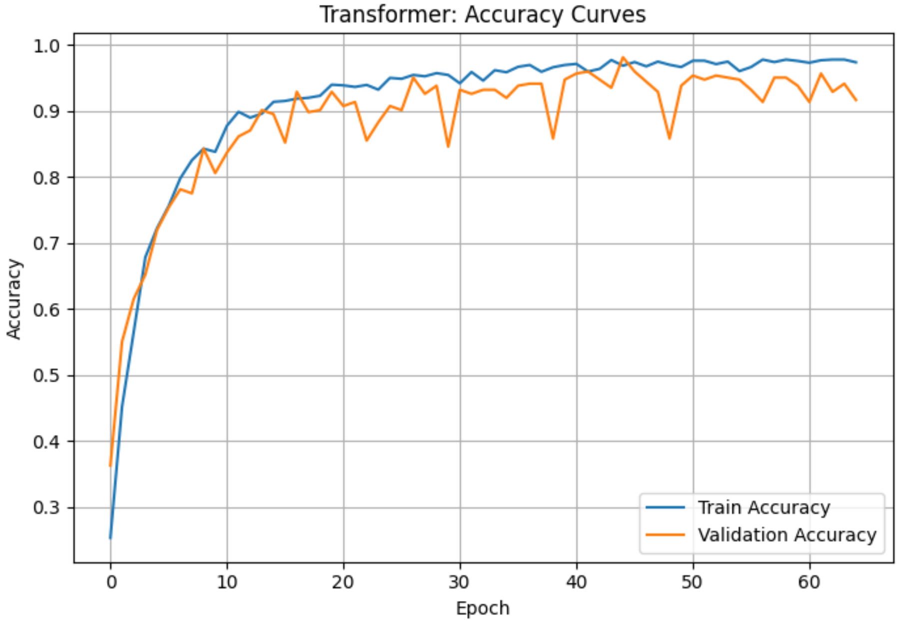
- Baseline LSTM model performed the worst as it processes information **linearly** and struggle with long dependencies
- BiLSTM performed better than the baseline as it reads sequence in **both directions** simultaneously
- Addition of 'Attention' allows model to focus on specific **relevant part** of input sequence
- Transformer achieved the highest performance as it uses **Self-Attention** to process entire sequence in **parallel** [3]

FINDINGS: Baseline LSTM vs. BiLSTM+Attention vs. Transformer

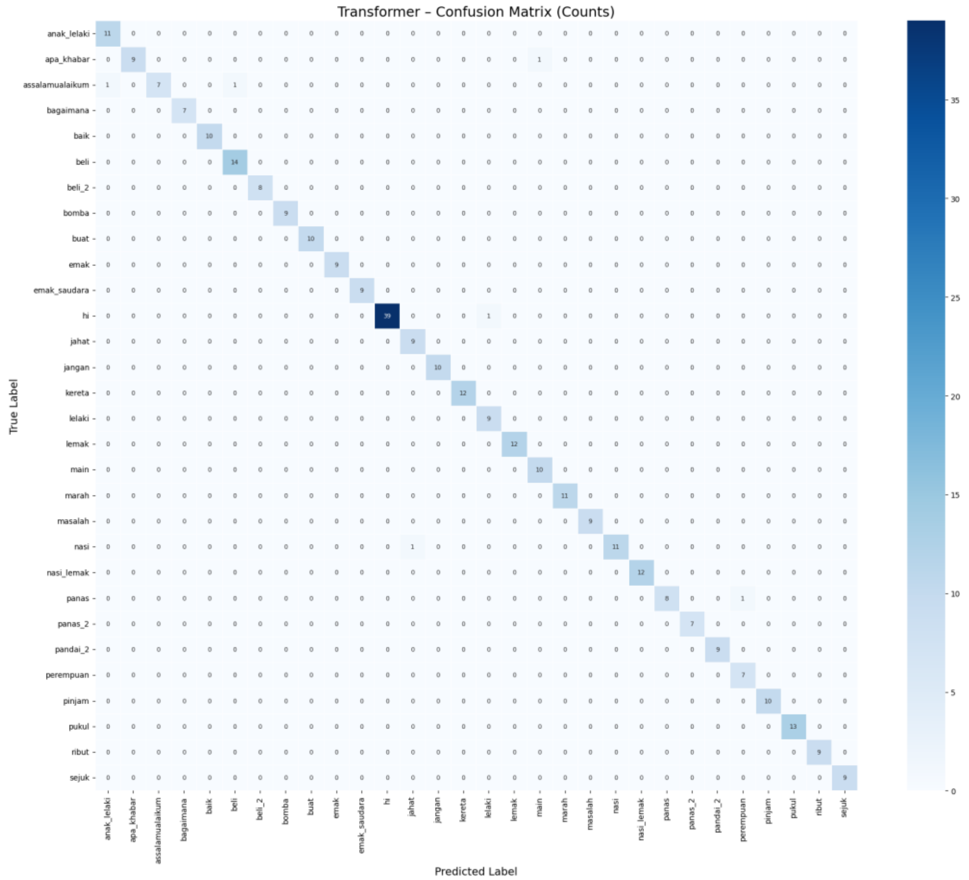


FINDINGS: Baseline LSTM vs. BiLSTM+Attention vs. Transformer

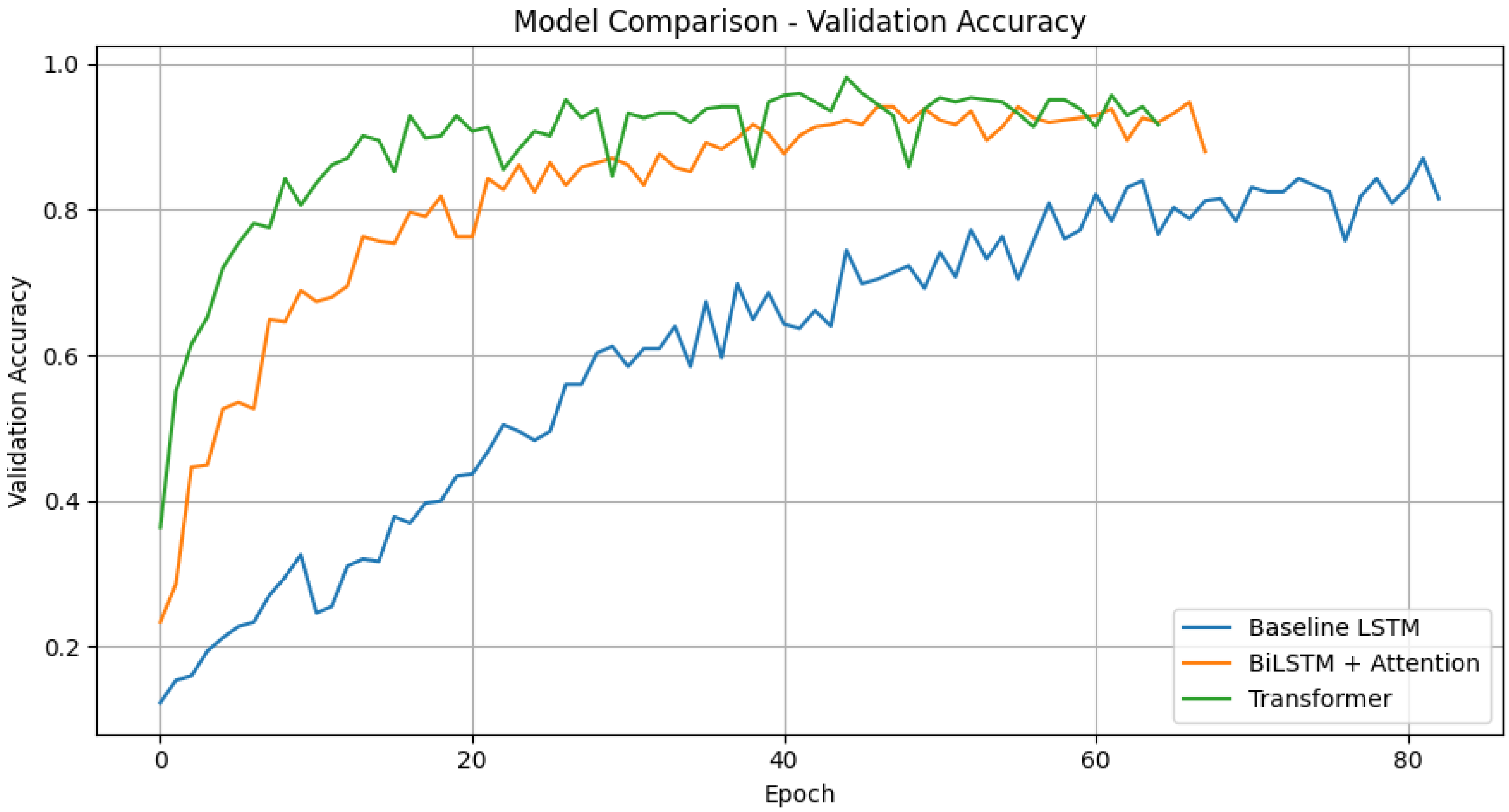




FINDINGS:
Baseline LSTM vs. BiLSTM+Attention vs.Transformer

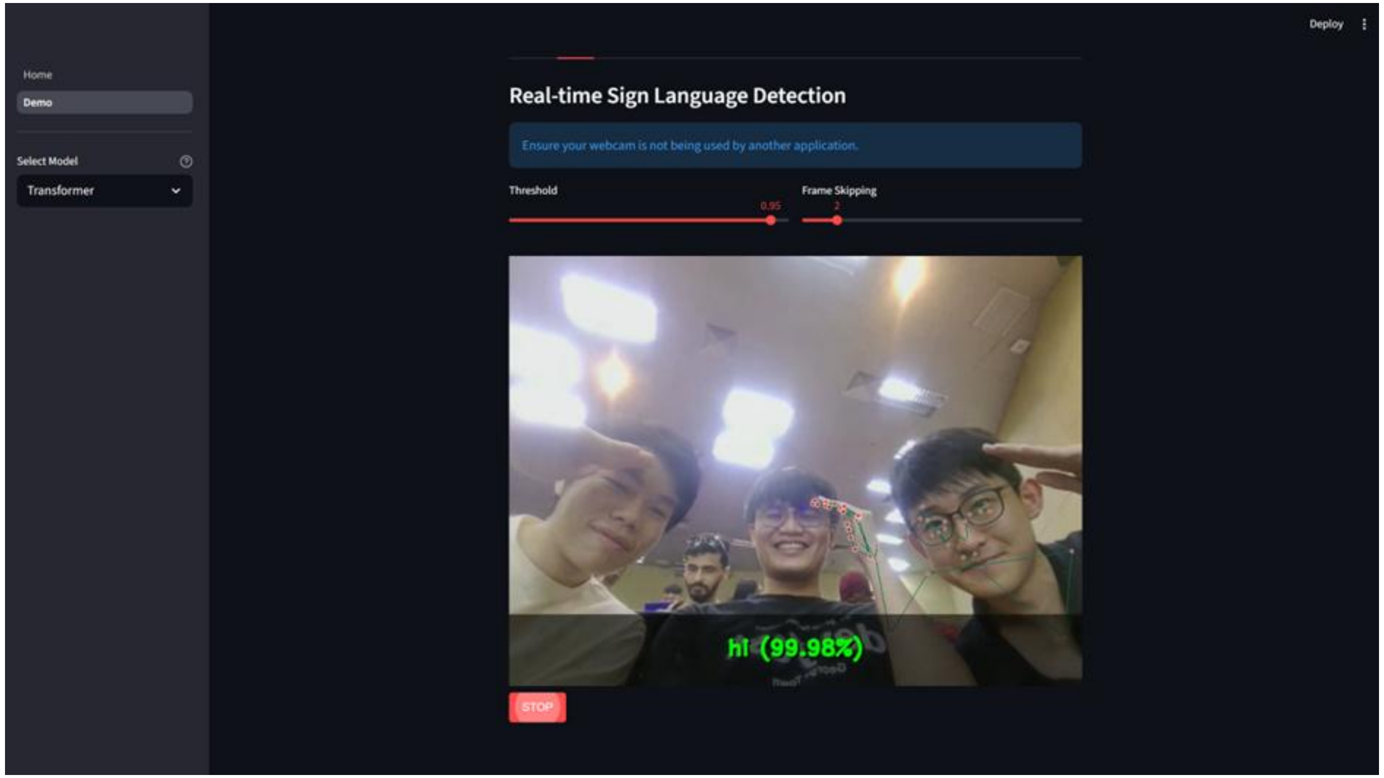


FINDINGS:
Baseline LSTM vs. BiLSTM+Attention vs.Transformer



UNIVERSITI MALAYA

DEMONSTRATION



- Demo Video link: https://drive.google.com/drive/folders/1WehNSv3otNHw_ztRMp-kk4gHAMXZHvt4?usp=drive_link
- Online Demo: https://isyaratai.streamlit.app/


anak_lelaki
bomba
_eeinko
beli
beli_2
assalamualaikum
baik
buat
emak
emak_saudara
hi
jahat
jangan
kereta
lelaki
lemak
main
pandai_2
marah
masalah
panas_2
nasi_lemak
panas
pinjam
nasi
perempuan
pukul
ribut
sejuk

DISCUSSION

- The prototype demonstrates **feasible recognition of isolated MSL glosses** using video-based landmark extraction and temporal modeling.
- Successful **integration into a simple application workflow** indicates practical usability in learning and demonstration settings.
- Performance is limited by small dataset size, class imbalance, signer variation, and sensitivity to environmental conditions.
- As a result, the system is currently best suited for controlled environments such as MSL learning, training and demonstrations.

CONCLUSION

- This project demonstrated the feasibility of using computer vision and deep learning for Malaysian Sign Language recognition
- Landmark-based feature extraction combined with an LSTM model was effective in capturing temporal gesture patterns
- Model performance was strongly influenced by class selection, data balance, and temporal sampling strategies
- Optimization techniques such as temporal consistency, stratification, and batch training significantly improved accuracy and generalization
- Limitations include a limited dataset size, isolated gesture recognition, and potential overfitting in small-class models
- Future work may include expanding the dataset, incorporating continuous sign recognition, and exploring more advanced temporal models

Q & A

REFERENCES

- [1] A. A. Chong, V. Yee, R. Bee, and M. Hussain, "Language Barriers in Deaf-Centred Classroom: Perspectives from Malaysian Deaf Adults," *Journal of Special Needs Education*, vol. 11, p. 2021, 2021.
- [2] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pp. 1448–1458, Mar. 2020, doi: 10.1109/WACV45572.2020.9093512.
- [3] L. T. Woods and Z. A. Rana, "Modelling Sign Language with Encoder-Only Transformers and Human Pose Estimation Keypoint Data," *Mathematics 2023, Vol. 11,* vol. 11, no. 9, May 2023, doi: 10.3390/MATH11092129.