

# WQF7006 COMPUTER VISION AND IMAGE PROCESSING

GROUP ASSIGNMENT

MALAYSIAN SIGN LANGUAGE TRANSLATION CASE STUDY

*Home of the Bright, Land of the Brave*  
*Di Sini Bermulanya Pintar, Tanah Tumpahnya Berani*



UNIVERSITI  
MALAYA

# GROUP INFORMATION

- OCC 1, GROUP 6
- GROUP MEMBERS:
  1. Chee Zhen Yu (24088354)
  2. Cheong Yi Fong (U2005327)
  3. Hong Jia Heng (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 Zhen Yu**  
**Software Developer**

Intergrated 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
- 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
- **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

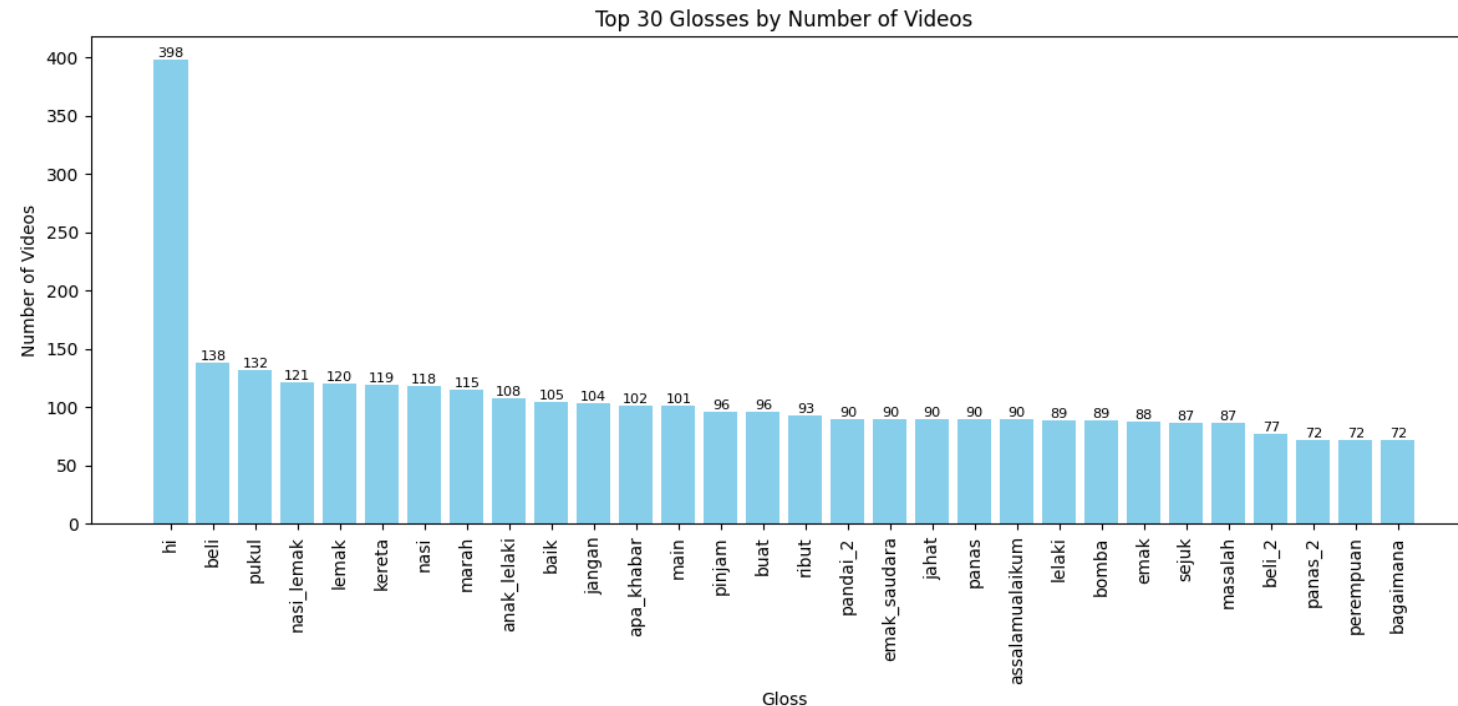


Figure shows top 30 selected glosses by number of videos:

- ['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 Holistic 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

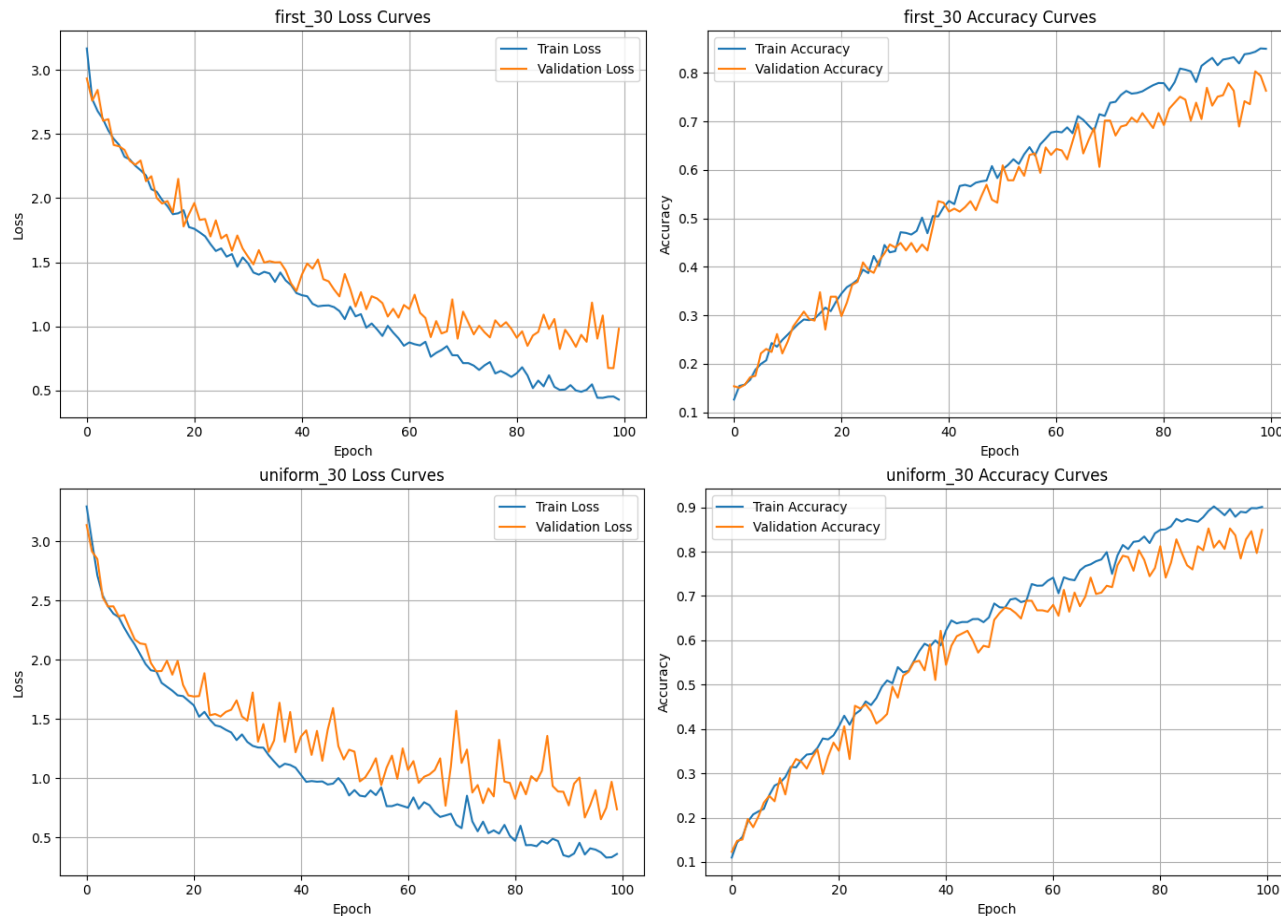
What we did?

- 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:

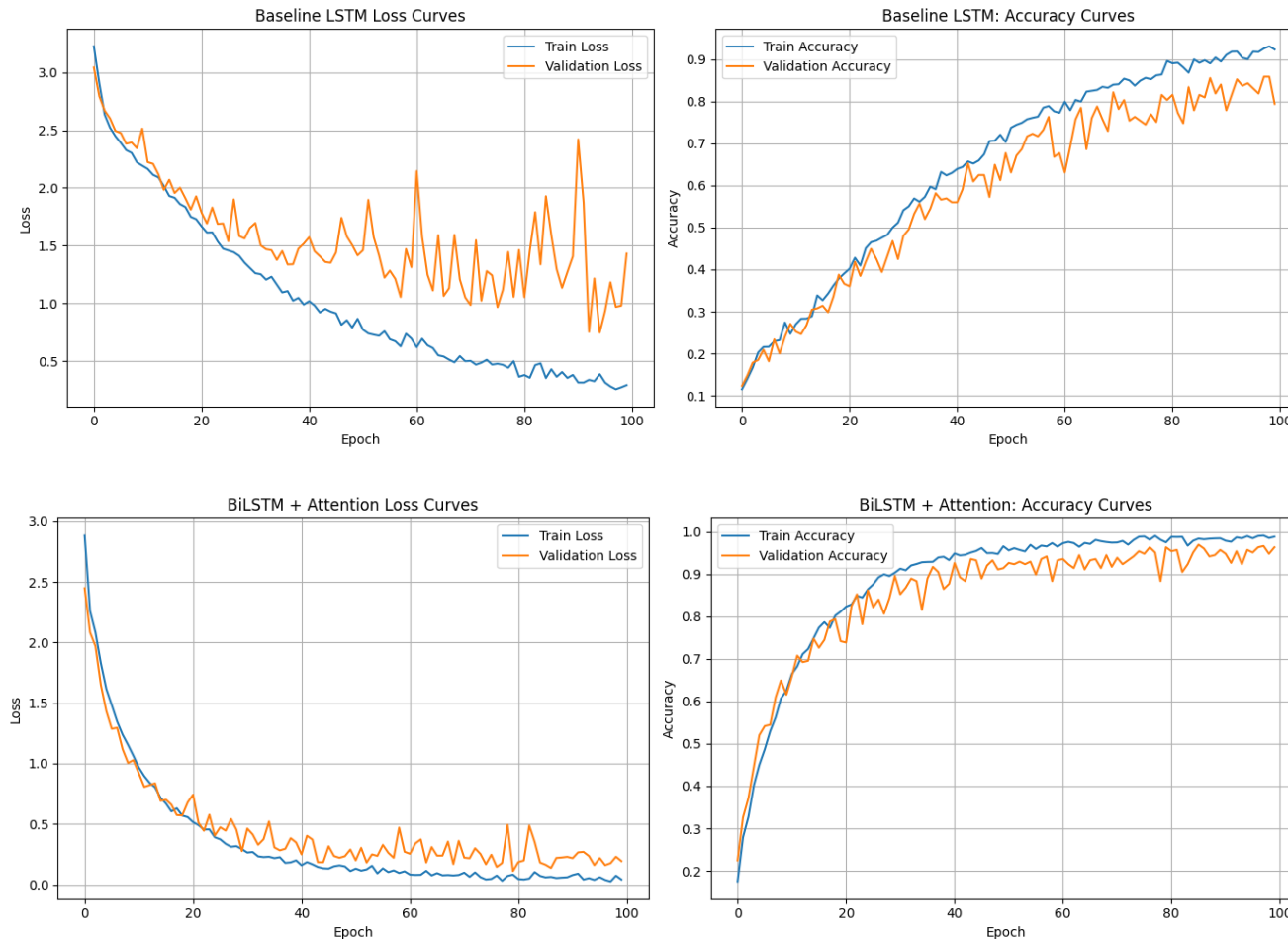
Strategy	Test Loss	Test Acc
First_30	0.9803	0.7631
Uniform_30	0.7391	0.8492

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

Model	Test Accuracy	Test F1 Score
Baseline LSTM	0.8431	0.84
BiLSTM + Attention	0.9631	0.96
Transformer	0.9662	0.97

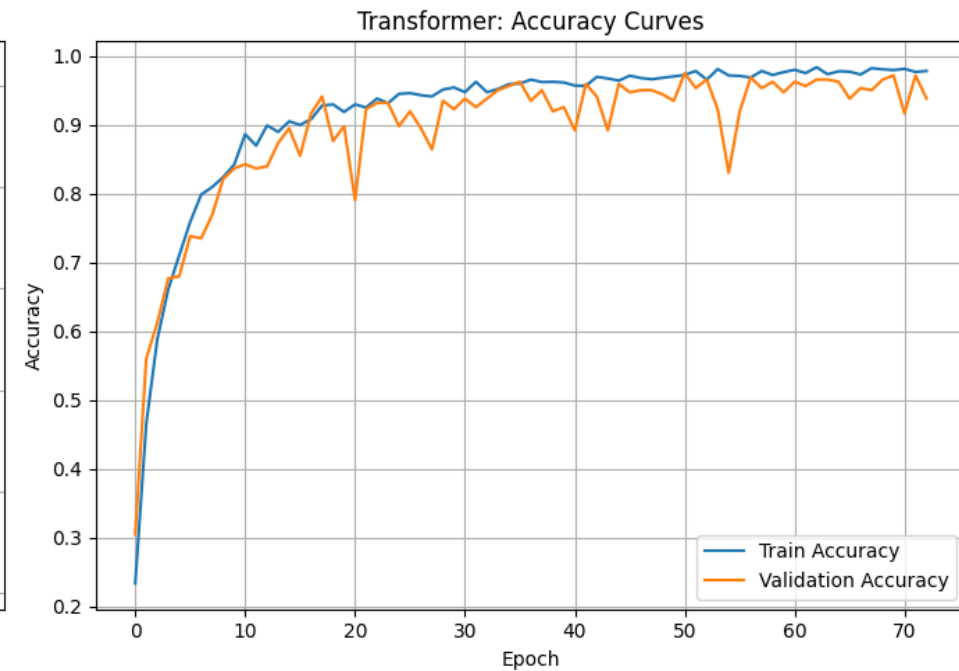
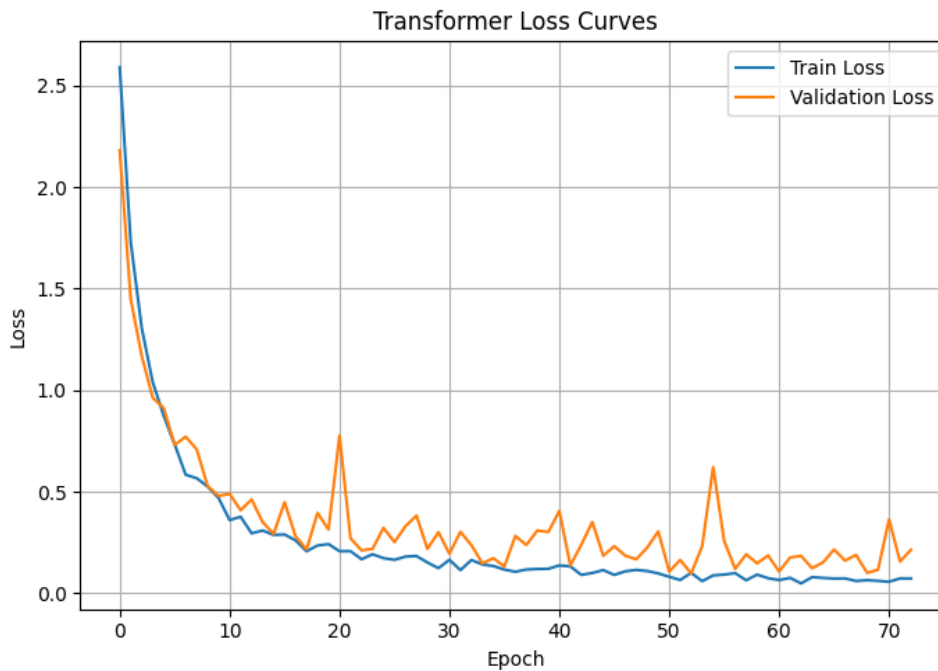
- Baseline LSTM model performed the worst as it processes information **linearly** and struggle with long dependencies
- BiLSTM performed better than the baseline as it reads sequence in **both directions** simultaneously
- Addition of 'Attention' allows model to focus on specific **relevant part** of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

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



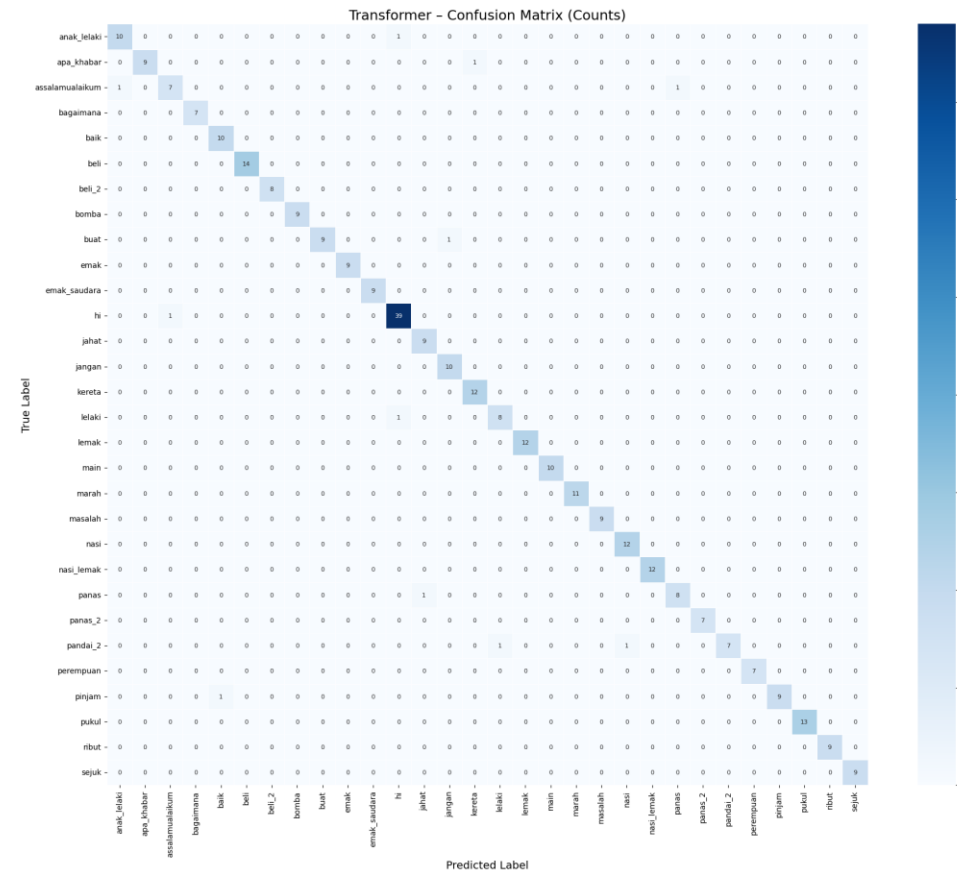


# FINDINGS: Baseline LSTM vs. BiLSTM+Attention vs. Transformer



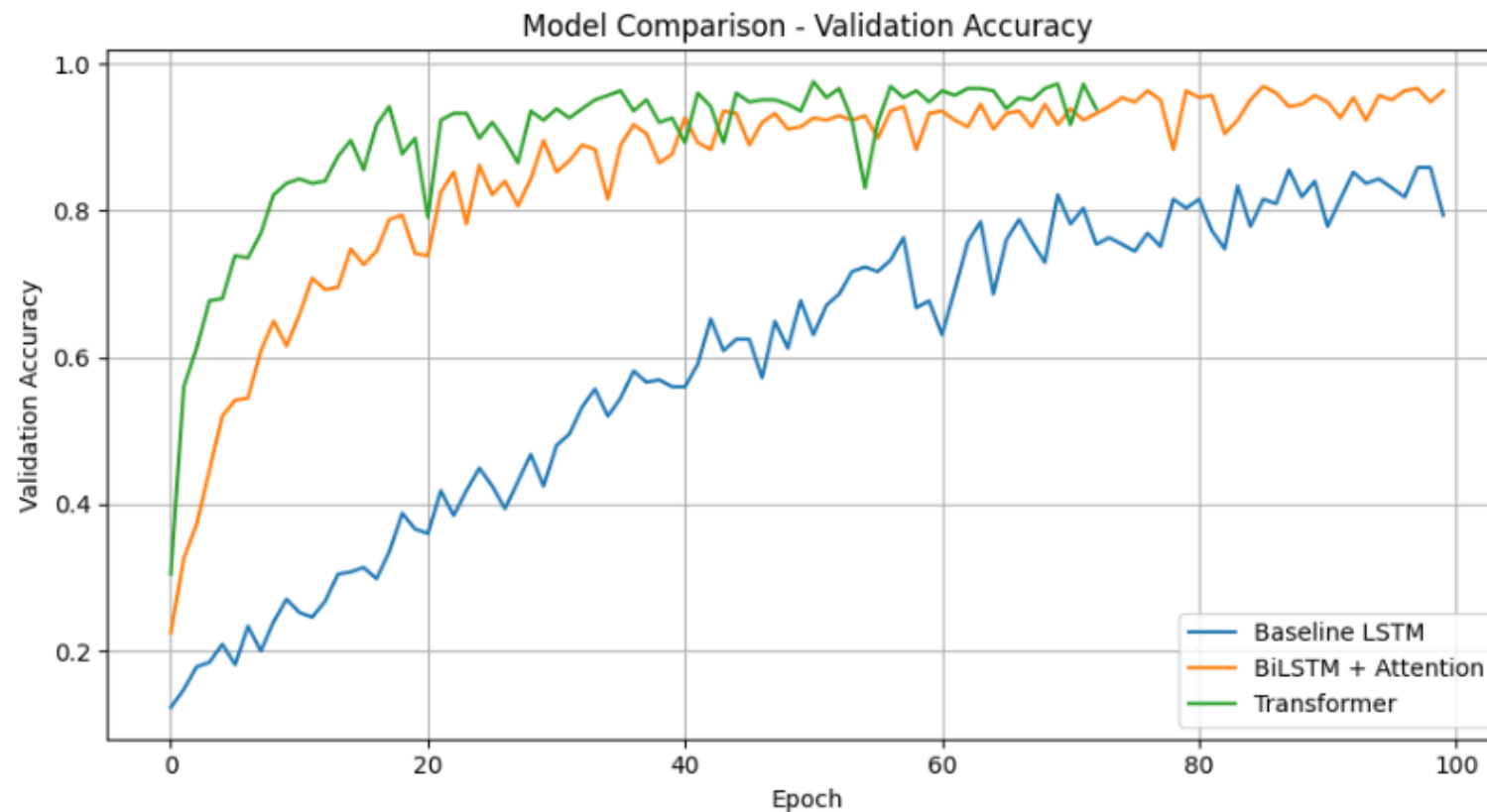
# FINDINGS:

## Baseline LSTM vs. BiLSTM+Attention vs. Transformer

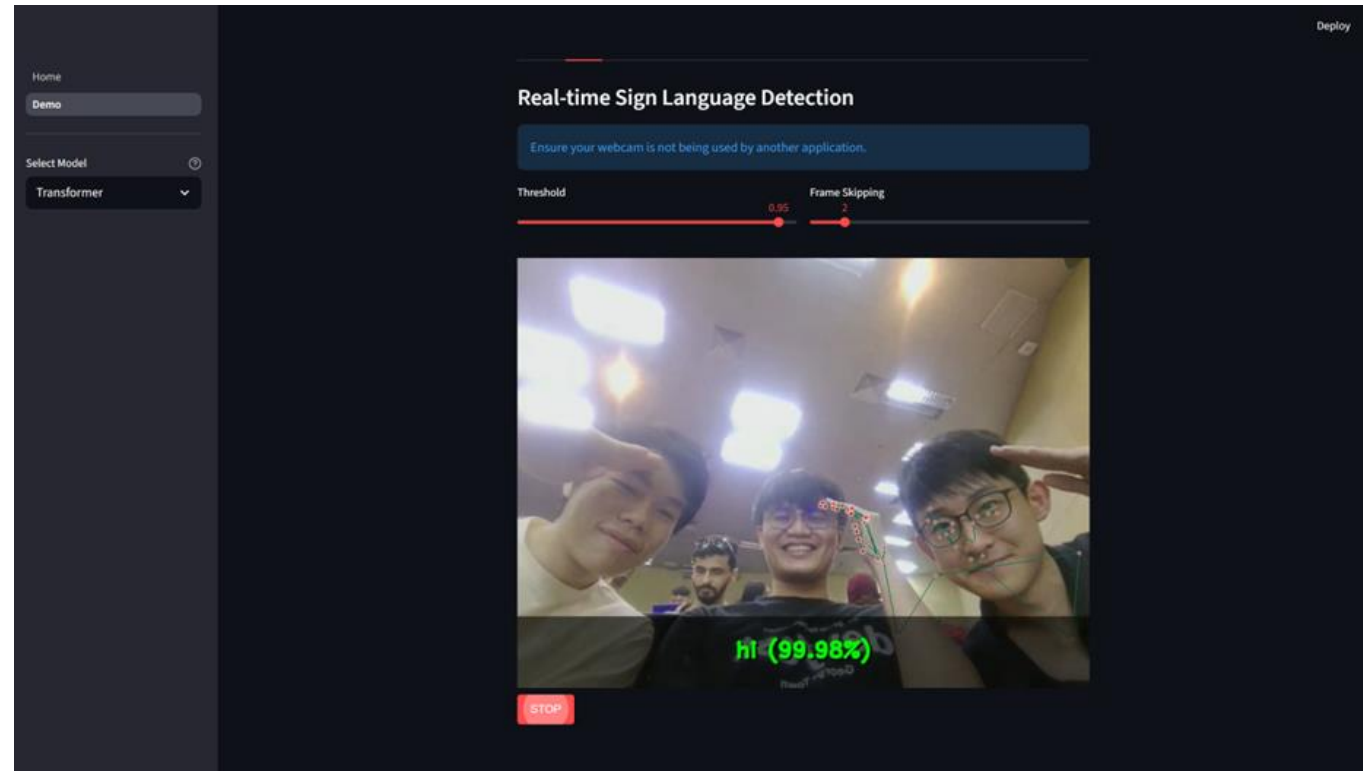


# FINDINGS:

## Baseline LSTM vs. BiLSTM+Attention vs. Transformer



# DEMONSTRATION



- Demo Video link: [https://drive.google.com/drive/folders/1WehNSv3otNHw\\_ztRMp-kk4gHAMXZHvt4?usp=drive\\_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  
assalamualaikumbaik buat  
emak emak\_saudara hi jahat  
jangan kereta lelaki lemak main  
pandai\_2 marah masalah  
panas\_2 nasi lemak panas  
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

*Home of the Bright, Land of the Brave*  
*Di Sini Bermulanya Pintar, Tanah Tumpahnya Berani*



[www.um.edu.my](http://www.um.edu.my)



[universityofmalaya](https://www.facebook.com/universityofmalaya)



[unimalaya](https://www.instagram.com/unimalaya)



[uniofmalaya](https://www.youtube.com/uniofmalaya)



UNIVERSITI  
MALAYA