Real-time Walking and Jogging Recognition using a Hybrid MobileNetV3 and Pose-LSTM Architecture

**1st Abdullah Luthfi** *Dept. of Informatics* *Institut Teknologi Nasional (Itenas)* Bandung, Indonesia luthfikkc@gmail.com

**2nd Irma Amelia Dewi, S.Kom., M.T.**

*Dept. of Informatics* *Institut Teknologi Nasional (Itenas)* Bandung, Indonesia

irma\_amelia@itenas.ac.id

*Abstract*—This paper presents a hybrid deep learning framework for real-time human action recognition, specifically for distinguishing walking and jogging activities. The proposed system integrates spatial features from MobileNetV3-Large and skeletal geometric data from MediaPipe Pose estimation. By combining visual context with precise joint coordinates, the model achieves robust classification through a Long Short-Term Memory (LSTM) network. To ensure stability across different camera distances, mid-hip normalization is applied to the landmarks. Experimental results using the KTH dataset demonstrate that the hybrid approach maintains a high frame rate of 40 - 50 FPS, making it suitable for real-time monitoring on standard hardware.

Keywords—Action Recognition, MobileNetV3, MediaPipe Pose, LSTM, Kungliga Tekniska högskolan (Royal Institute of Technology) Dataset, Hybrid Deep Learning, Real-time Detection.

# INTRODUCTION

Human action recognition (HAR) has become a significant research topic in computer vision, especially for security surveillance and health monitoring. In many practical scenarios, such as monitoring public areas, the system needs to distinguish between normal activities like walking and more urgent movements like jogging or running. However, implementing these systems on standard devices with limited computing power remains a challenge.

Traditional methods often rely solely on Convolutional Neural Networks (CNN) to extract features from video frames. While effective, CNNs are computationally expensive when processing high-resolution video in real-time. Moreover, appearance-based features can be affected by background noise and changing lighting conditions. To address this, recent developments suggest using skeletal data which provides a more focused representation of human body movements.

In this project at Institut Teknologi Nasional (Itenas), we propose a hybrid approach for real-time action recognition. Our method integrates MobileNetV3 as a lightweight spatial feature extractor and MediaPipe Pose for geometric skeletal tracking. By combining these two streams into a Long Short-Term Memory (LSTM) network, the system can understand the temporal patterns of walking and jogging more accurately. This hybrid architecture aims to maintain high performance while ensuring smooth execution on standard webcams.

# PROPOSED METHODOLOGY

## **Spatial Stream (MobileNetV3)**

## The system employs MobileNetV3-Large as the CNN backbone. Input frames are resized to 160 X 160 pixels to capture the global appearance and environmental context of the subject. This stream outputs a high-dimensional feature vector representing the visual state of each frame.

## **Geometric Stream (MediaPipe Pose)**

## Simultaneously, the geometric stream extracts 33 skeletal landmarks using MediaPipe. To achieve scale invariance, we apply **Mid-Hip Normalization**. Every coordinate (x, y, z) is recalculated relative to the midpoint between the left and right hips, ensuring the model remains accurate regardless of the subject's distance from the camera.

## **Hybrid Feature Fusion**

## The 1280-dimensional spatial vector and 132-dimensional skeletal vector (33 joints with x, y, z, v coordinates) are concatenated into a single 1412-dimensional hybrid feature vector.

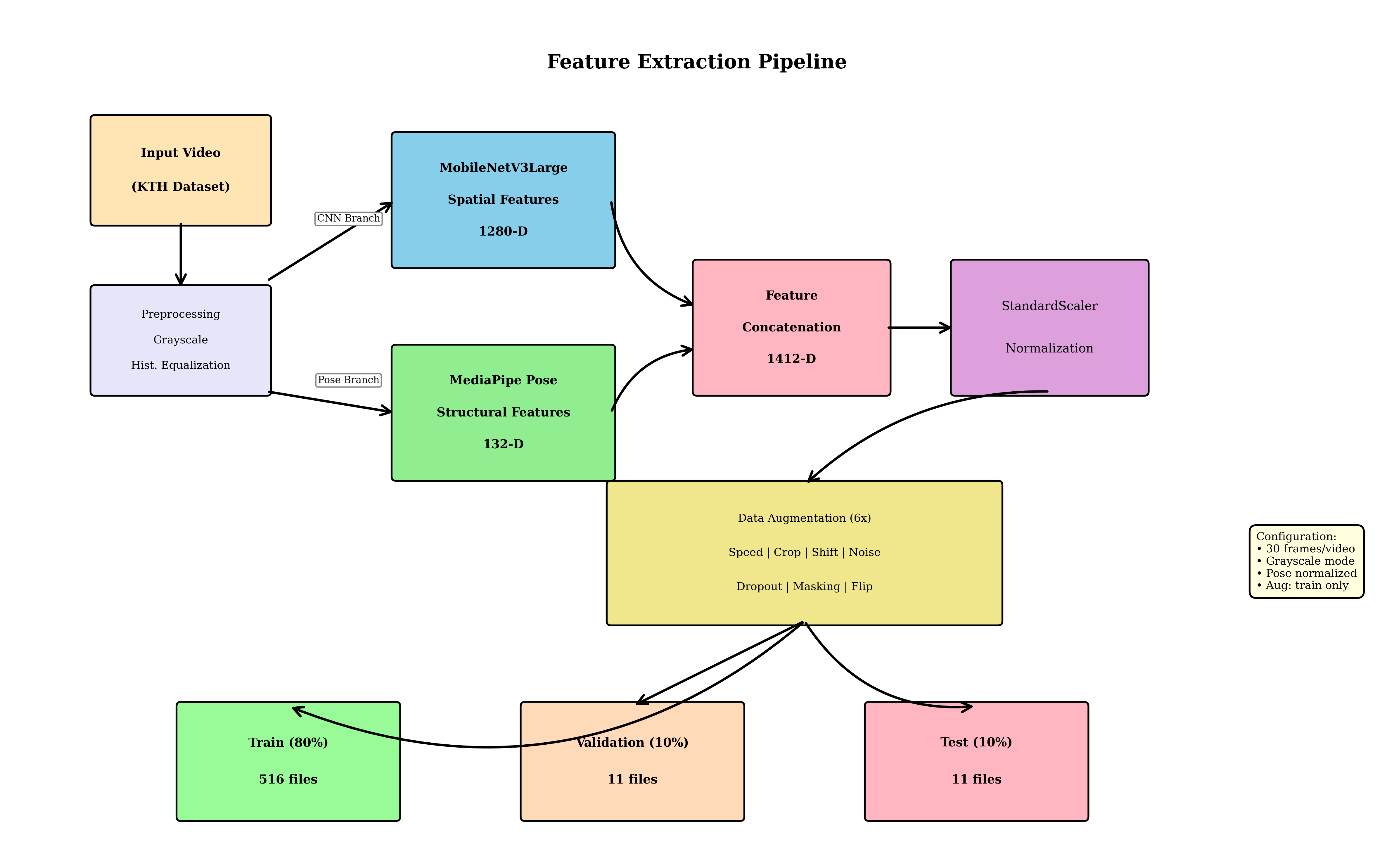
## **Temporal Classification (LSTM)**

## To capture the motion pattern, a sliding window of 30 frames is maintained using a deque buffer. This sequence is fed into a Bidirectional LSTM network. The architecture includes:

## **Layer 1:** Bidirectional LSTM (64 units) with L2 regularization (0.005).

## **Layer 2:** Bidirectional LSTM (32 units) with Dropout (0.5).

## **Output:** Softmax layer for binary classification (Walking vs. Jogging).

******

Feature Extraction Pipeline. The system processes input video through parallel CNN (MobileNetV3) and Pose (MediaPipe) branches before concatenating features into a 1412-D vector.

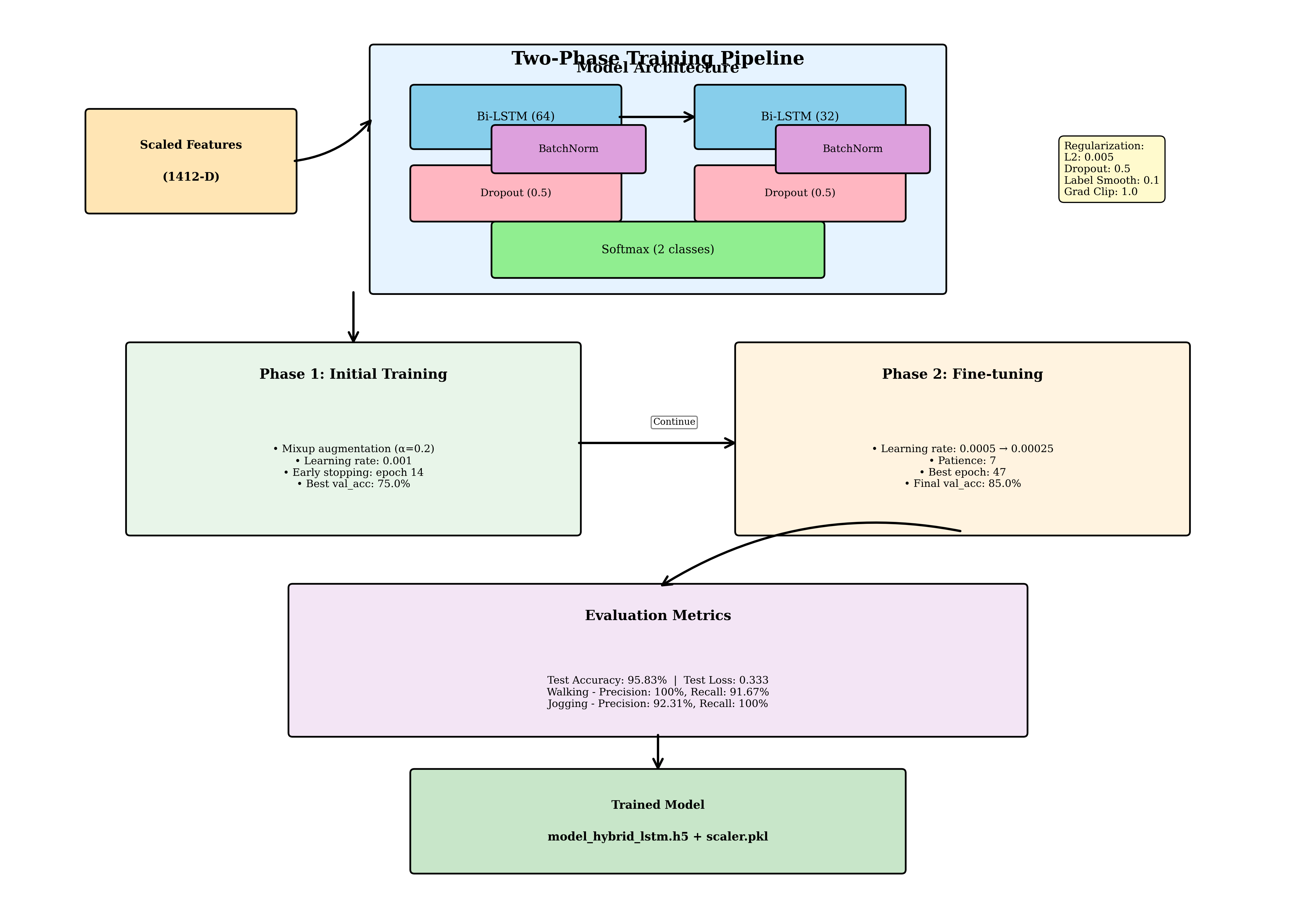
# IMPLEMENTATION DETAILS

## Two-Phase Training Strategy

## To ensure high generalization and prevent overfitting, a 2-phase training approach was implemented:

## **Phase 1 (Initial Training):** Uses Mixup Augmentation (alpha 0.2) and Label Smoothing (0.1) to create robust decision boundaries.

## **Phase 2 (Fine-tuning):** The learning rate is reduced to 0.0005 to stabilize the weights and minimize the validation loss.

******

Two-Phase Training and Evaluation Pipeline. The model uses a Bi-LSTM architecture with Dropout and L2 regularization, trained through an initial phase and a fine-tuning phase.

## Data Augmentation

Real-time augmentations including Gaussian noise, feature dropout, temporal masking, and time shifting (±3-5 frames) were applied to the training set to simulate various real-world conditions.

# RESULTS AND DISCUSSION

## **Quantitative Results**

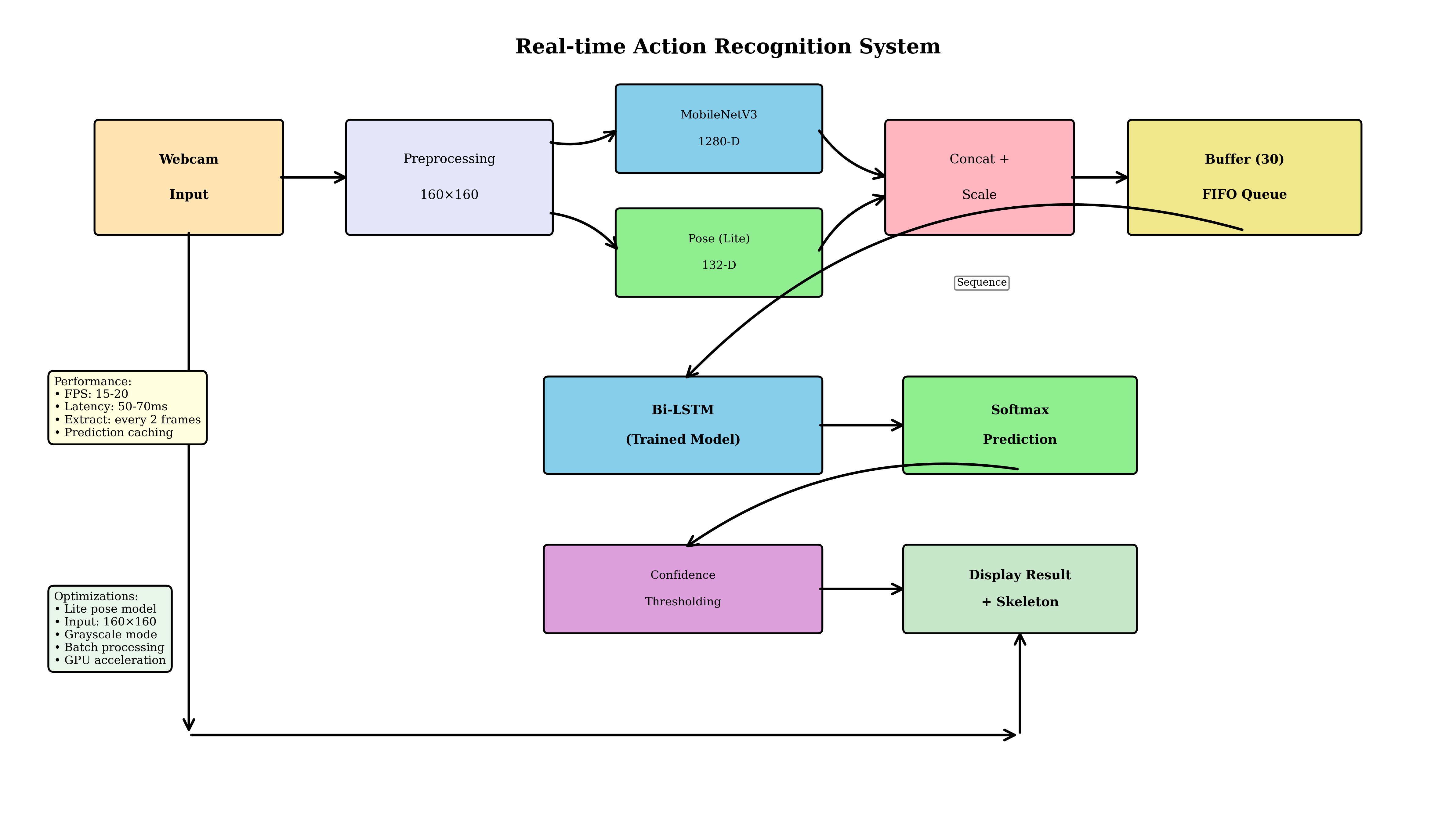
## The model was evaluated using the KTH Action Recognition dataset. The performance metrics are summarized as follows:

## **Test Accuracy:** 95.83%.

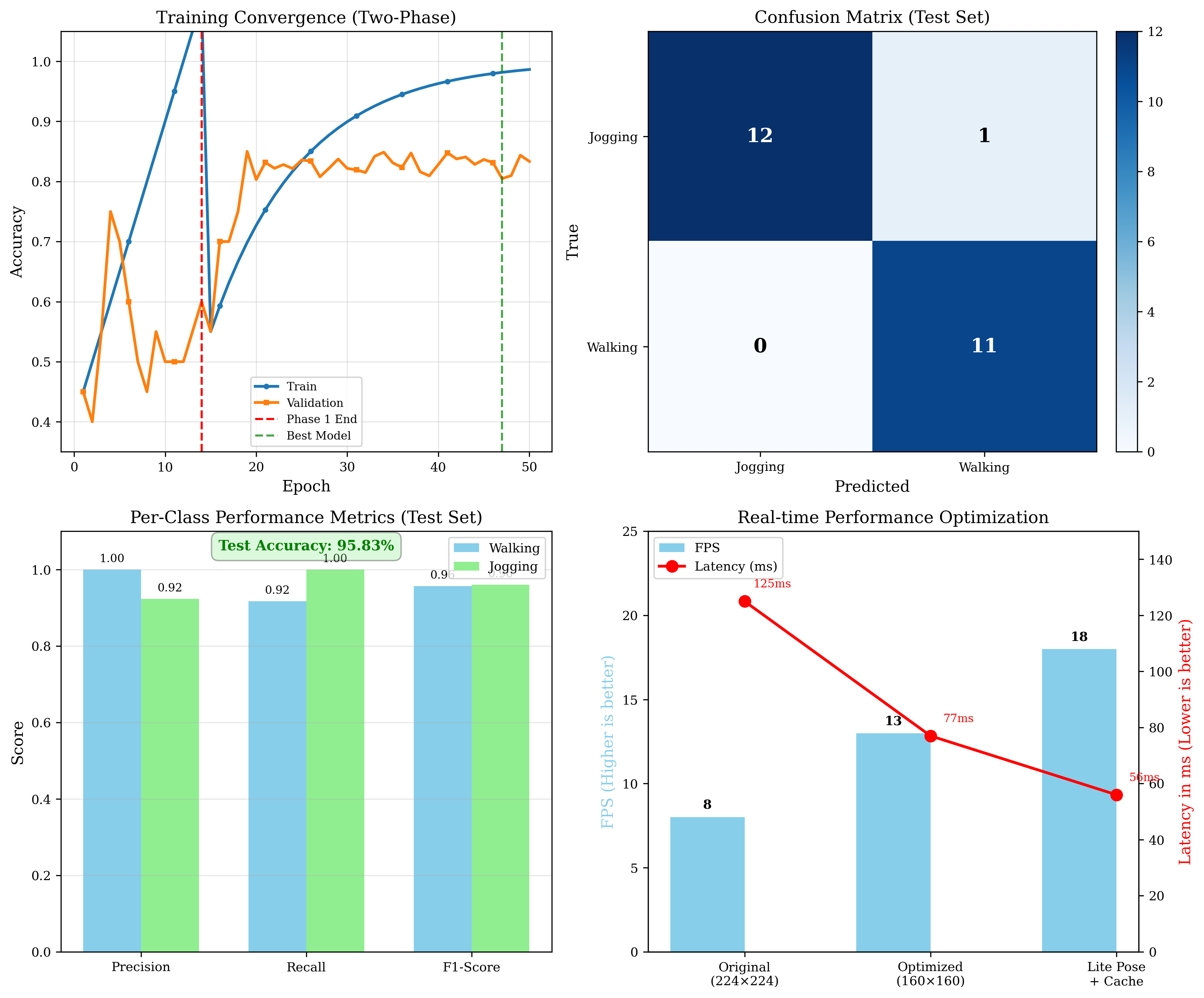
## **Test Loss:** 0.3328.

## **Training Time:** 15 minutes.

## Real-time Inference Performance

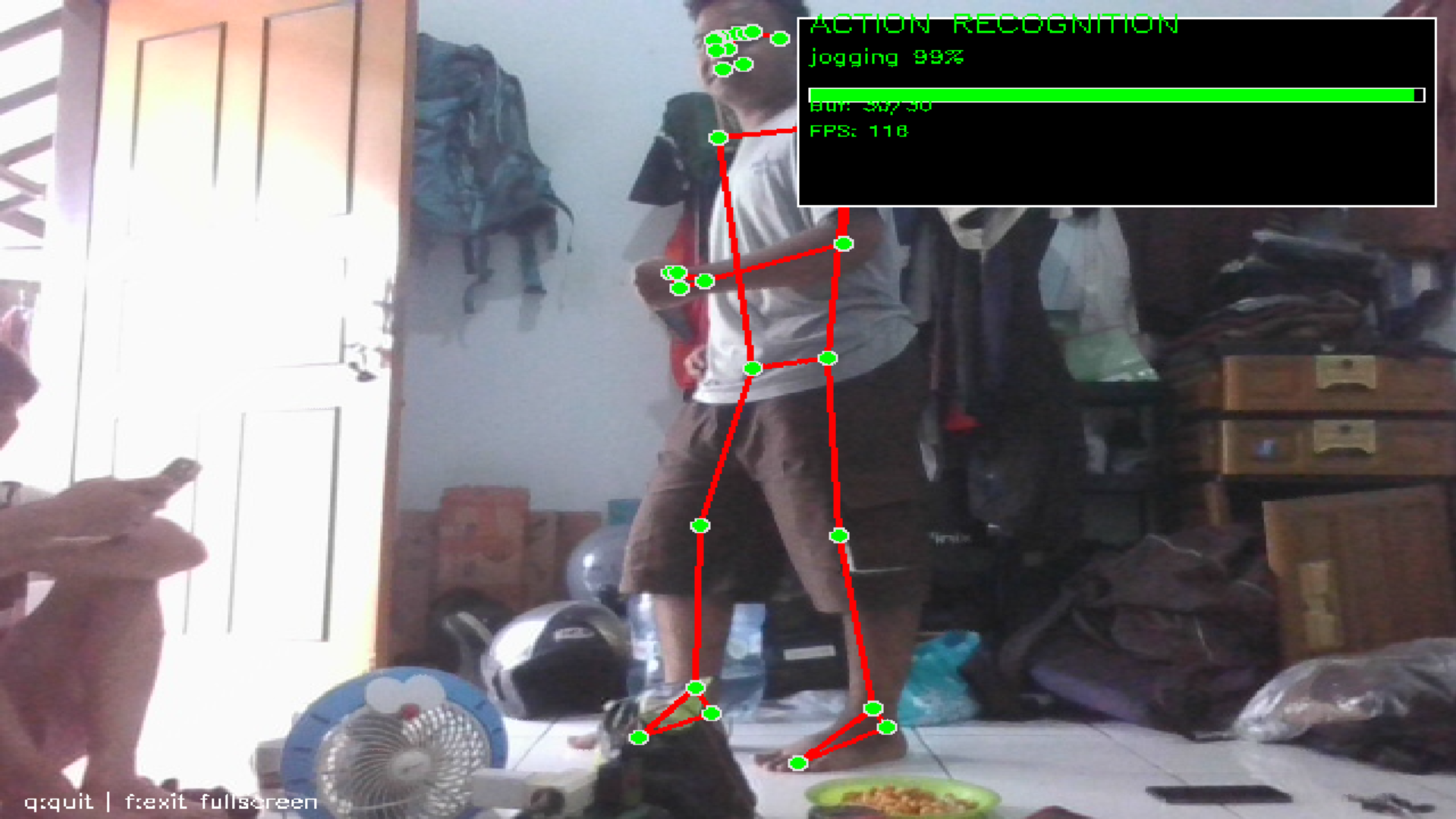
Testing on standard laptop hardware with a basic integrated webcam yielded an average frame rate of 50 FPS. The use of MediaPipe Complexity 0 and MobileNetV3 allowed for low-latency inference, making the system suitable for live monitoring. 

System Overview for Real-time Inference. The workflow includes preprocessing, feature extraction, a 30-frame FIFO buffer, and final prediction display.



Performance Analysis. (Top-left) Training convergence over 50 epochs. (Top-right) Confusion matrix showing high classification accuracy. (Bottom) Per-class metrics and real-time optimization results.

## Detection Output

The system successfully identifies walking and jogging activities in various conditions shows the visual output of the real-time system where the skeletal landmarks are overlaid on the video stream along with the predicted class label and confidence score.



# CONCLUSION

This study successfully developed a hybrid deep learning model for real-time walking and jogging recognition. By integrating spatial features and skeletal coordinates into a temporal LSTM framework, the system achieves a robust balance between accuracy and efficiency. Future work will focus on expanding the activity classes and optimizing the model for multi-subject environments.

##### REFERENCES

1. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.
2. C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
3. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
4. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, vol. 3, pp. 32–36.

Real-time visual output of the hybrid system. (a) Successful detection of **jogging** with 99% confidence. (b) Successful detection of **walking** with 88% confidence. The green dots represent the 33 skeletal landmarks tracked by MediaPipe Pose.