MedAssist: An Automated Solution for The Assessment of Medication Intake

Adem Kaya, George Lalidis, Giedrius Mirklys, Tam Van

21-06-2025

1 Introduction

Within the scope of the AI in the Professional Workfield course (SOW-MKI76), the task was to develop a project for a company selected from the Masters Challenge platform. Based on a collective interest in societal impact and healthcare, MedAssist was chosen as the focus of this project.

1.1 MedAssist

MedAssist specializes in the development of medication dispensary devices. These devices incorporate automatic medication release to assist patients in adhering to their prescribed schedules. Additionally, they are equipped with integrated cameras capable of recording videos of patients positioned directly in front of the device.

1.2 The Problem

Currently, the video data acquired by these devices undergoes manual review to ascertain whether patients have successfully taken their medication. The primary challenge lies in the time-intensive nature of this manual process. To address this, MedAssist seeks an automated solution, specifically an AI system capable of accurately assessing medication intake.

1.3 The Goal

The objective is to develop an AI system that can automatically assess medication adherence with high accuracy. Emphasis will be placed not only on maximizing overall accuracy but also on minimizing false positive predictions, where the AI incorrectly indicates medication intake, which is particularly critical in healthcare settings.

2 Methods

2.1 Data Labelling and Processing

MedAssist provided a collection of videos filmed using their medication tracking devices. The initial dataset consisted of 85 unlabeled videos, exhibiting variations in length, subjects, surroundings, and lighting conditions. Following manual annotation, 71 videos

were identified as depicting medication intake. It is important to note that 20 videos presented challenges in determining medication intake; however, actions resembling eating or bringing an item close to the mouth were conservatively labeled as positive instances.

A supplementary dataset of 29 videos (23 positive instances) was subsequently provided, featuring increased subject variation. The combined dataset comprised 114 videos with an 82% class imbalance. This dataset was partitioned into training (70%), testing (20%), and validation (10%) sets.

Each video, approximately 30 seconds in duration with a frame rate of 60 fps (approximately 1800 frames), was processed by extracting 10 frames at uniform intervals due to computational constraints. These frames were stored as a NumPy zipped array and preprocessed using the HAR model from Hugging Face[1] during both training and testing.

Given that the fine-tuned model classifies individual frames rather than entire videos, a method for efficient video processing was required. To achieve this, 10 frames were extracted from each video at uniform intervals, enabling processing of the entire video while analyzing only a subset of frames. This approach assumes that for videos annotated with medication intake, at least one extracted frame will capture the action, allowing the model to identify the activity while minimizing computational cost.

2.2 Human Activity Detection (HAD)

Human activity detection (HAD) is a pretrained transformer that focuses on recognizing and classifying human activities from different possible data modalities. This model could detect multiple activities in a single frame, making it suitable for the task of medication intake assessment. The model was trained on a diverse dataset of human activities, enabling it to generalize well across various scenarios. For this project, the model was fine-tuned to specifically recognize medication intake activities:

I. Medication intake: The subject has taken their medication at some point in the video.

II. No medication intake: The subject has **not** taken their medication at any point in the video.

2.2.1 SligLIP 2

Due to the dataset's limited size, a pre-trained model based on the SligLIP 2 transformer architecture was utilized instead of training a HAD model from scratch. The implementation of this model, made publicly available and accessible through the HuggingFace model hub by Prithiv Sakthi, has demonstrated strong performance in detecting and classifying 15 distinct human activities in still images.

Although the model was not explicitly trained for medication intake assessment, its capabilities in recognizing *drinking* and *eating* activities were deemed particularly relevant, with F1 scores of 0.89 and 0.93, respectively. For a comprehensive overview of the model's performance across all 15 activities, refer to the confusion matrix presented in Figure 1.

[1] `https://huggingface.co/Adekiii/HAR-medication-finetuned_v2/tree/main`

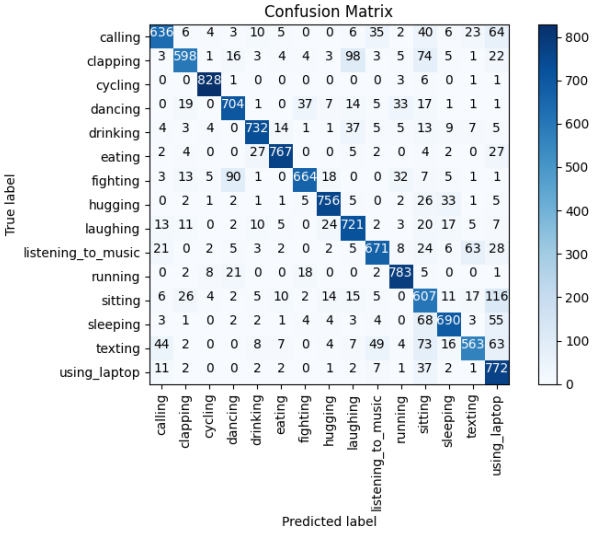

Figure 1: SligLIP 2 Confusion Matrix for Activity Classification

2.2.2 Fine-tuning

Since the pre-trained model was not specifically trained for medication intake, fine-tuning was performed. This involved manual annotation of individual frames within the training set video.

3 Results

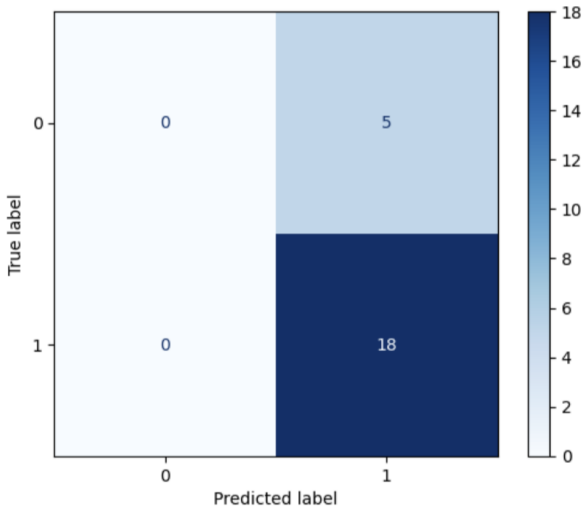

Figure 2: SligLIP 2 Confusion Matrix for Activity Classification

4 Conclusion

Summarize the results and discuss the implications of your findings.

5 Discussion and Recommendations

References