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 for 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, accompanied by a nationwide shortage of caregivers. 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 of these

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, after discussion with the representative of MedAssist, 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.

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`

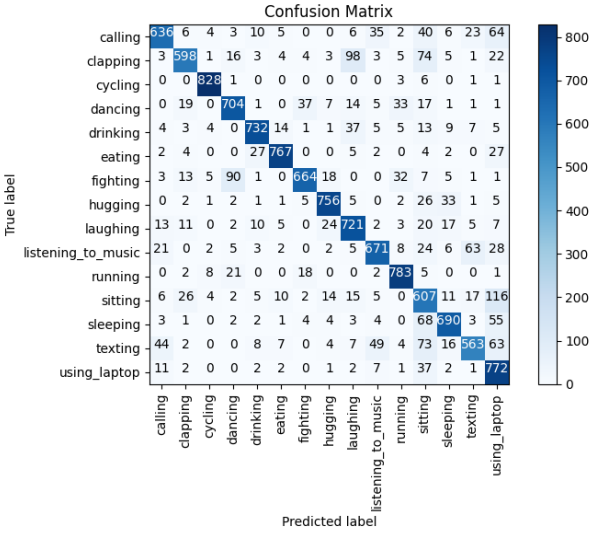

Figure 1: SligLIP 2 Confusion Matrix for Activity Classification

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. In total, 850 frames were manually annotated. Of these 850 frames, approximately 19% was labeled as positive cases of medication intake.

3 Results

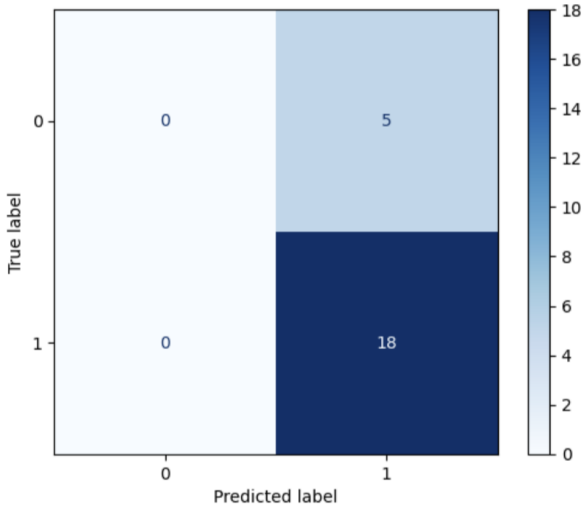

Figure 2: SligLIP 2 Confusion Matrix for Activity Classification

Epoch	Accuracy	Precision	Recall	F1
1	0.16	0.00	0.00	0.00
2	0.83	0.83	1.00	0.91
3	0.83	0.83	1.00	0.91

Table 1: Validation results per epoch

These results show how the model achieves a high F1 score, indicating a good balance between precision and recall.

4 Discussion

The model's ability to identify all cases involving medication intake might suggest that its performance is more than ideal. However, the fact that model always predicts the positive class would raise reasonable doubts about its true effectiveness. This is evidenced by the fact that all evaluation metrics involved–except recall which remains stable at 1.0 regardless the class distribution–depend solely on the number of cases in test set where medication was taken. Therefore, if the test set contains fewer cases of medication intake, these metrics would decrease accordingly. That said, although F1-score may appear high, it is not truly representative of the model's performance, as it also depends on the number of positive cases.

This unpredictable performance, along with the unexpected results, could stem from several factors. A large and diverse dataset is always essential for deep learning tasks, and this is especially critical in the healthcare domain. Including samples from the same subjects is legitimate in this context, as it reflects real-world scenarios where patients may have multiple medication events over time. However, given the small dataset size, this practice may contribute to suboptimal performance, leading to bias toward always detecting medication intake. Additionally, class imbalance–specifically, the oversampling of the positive class relative to the negative–must be considered as a potential contributing factor. Based on the provided labeling criteria, videos containing any motion toward the subject's mouth were to be classified as positive–a pattern the model successfully learned–even in cases where no actual medication was taken.

The pretrained model may also not be well-suited for this specific task. While it has successfully learned to recognize activities involving motion–particularly gesture toward the mouth–the specific task involves more nuanced actions. For example, the model cannot determine whether the patient actually swallowed the pill or whether the pill was present at all. As a result, any motion toward the mouth might be interpreted as medication intake, showcasing the model's limited ability to generalize effectively to the specific requirements of the task. Last but not least, the pretrained model used is computationally heavy, making it unsuitable for deployment on the dispenser. This indicates that an alternative approach should be considered–specifically, the implementation of a more lightweighted model.

5 Conclusion

The goal of this project was to develop an AI model that could be embedded in a dispenser equipped with a camera to detect whether patients have taken their medication. A

pretrained model was employed, originally trained on images depicting various activities, such as eating, walking, and smiling. Since the model was designed for image classification rather than video analysis, 10 frames were extracted from each video to adapt it to the task. The model effectively captured motion-related features, which justified its selection and integration into the system. It was subsequently finetuned for the specific task of medication intake detection. However, the results were not as promising as expected, and several factors may explain this outcome. The dataset was neither large nor diverse, containing limited samples, many of which involved similar individuals. This limitation introduced class imbalance, causing the model to become biased toward the positive class, likely contributing to its tendency to classify every sample as positive case. Finally, the pretrained model may ultimately not be ideal for this task, suggesting that a different approach should be considered–especially given that such a computationally heavy model cannot be embedded in the dispenser.

References