SIT225 Data Capture Technologies

Pass Task: Use case design

Name: **Hoang Long Tran**

ID: **s223128143**

Date Last Edited: September 15, 2024

Contents

1 Identifying a problem

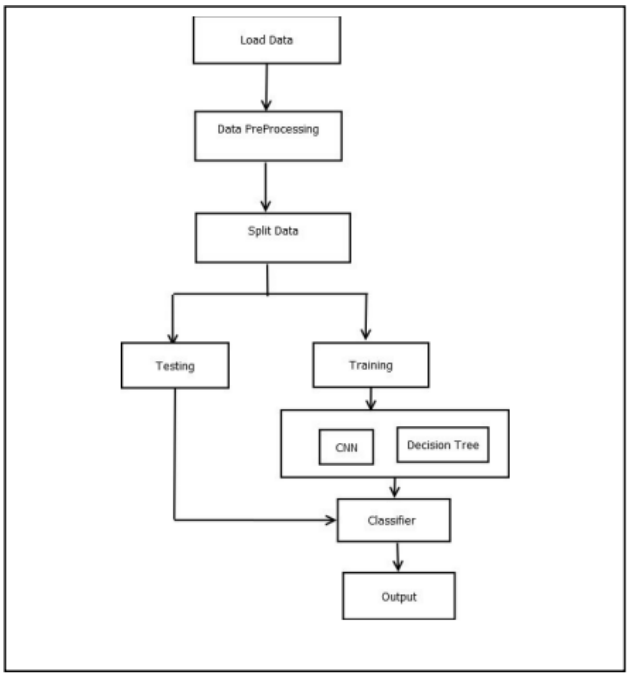
I have seen a couple of videos on how deaf people having a difficult time ordering food, especially at drive-through ordering. I want to apply my data skills into creating a way where those who have trouble hearing can have an easier time ordering their food.

2 Literature review

2.1 Article from Shaik et al. (2023)

This article proposes a design on hand gesture-based ordering system for businesses like restaurants. This design can solve problems like preventing the spread of germs, tackling language barriers and communication difficulties so people can order their food with just a simple wave, swipe, and other gestures. Note only that, implementing gesture-based system can help businesses on understanding the customers more by analyzing the most frequently used gestures and popular menu items, restaurants can optimize their menu, marketing and overall business.

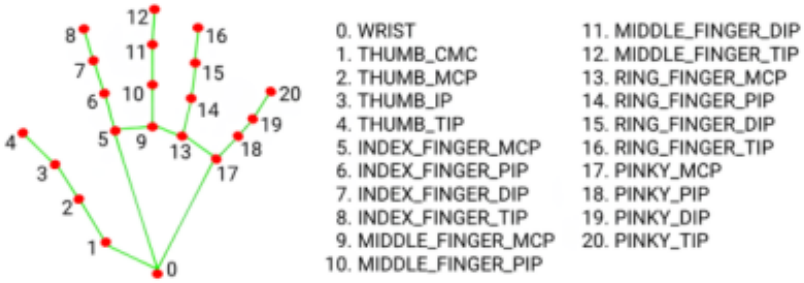
The article proposes this architecture to train their model.



2.1.1 Data collection

They collected data from many open repositories and extracted x and y coordinates of 21 specific hand keypoints using “MediaPipe” and “OpenCV”. They used “Palm Detection Model” from “MediaPipe” to

identify bounding boxes around rigid section like palms and closed fists, instead of detecting the whole hand due the different hand sizes and poses. Then they used "Hand Landmark Model" to identify 21 specific 3D coordinates (x, y, z) to map landmarks on the detected palm region.



Using the two models in MediaPipe, they process datasets like the American Sign Languages (ASL) and saved 21 points for each alphabet image. The alphabet images retain only and x and y coordinate, and remove the z.

2.1.2 Data cleaning

They removed null data and unclear images. Then normalized the data. Then split the data for 80% training and 20% testing. It is also mentioned to convert the image from an RGB color to HSV (Hue, Saturation, Value) or HSL where the picture is visually nicer to see, and some models may perform better.

2.1.3 Hybrid Machine Learning prediction via CNN and Decision tree

CNN is good at handling spatial data; Decision Tree is good in managing structured data related to food ordering system. By combining these two, we will have a nice predictive model.

CNN is implemented first to learn spatial hierarchies of features from the images. This CNN is different from regular ones, instead of moving the output layer, that layer will be used as input for the Decision Tree. Here is the outline of the hybrid model architecture:

CNN:

1. Input layer: input normalized x, y.
2. Convolutional layers: multiple layers tasked with extracting essential features from the data.
3. Pooling layers: reduce spatial dimensions (dimension reduction).
4. Fully Connected Layers: flatten the data and bridge the extracted data from CNN to Decision Tree

Decision Tree:

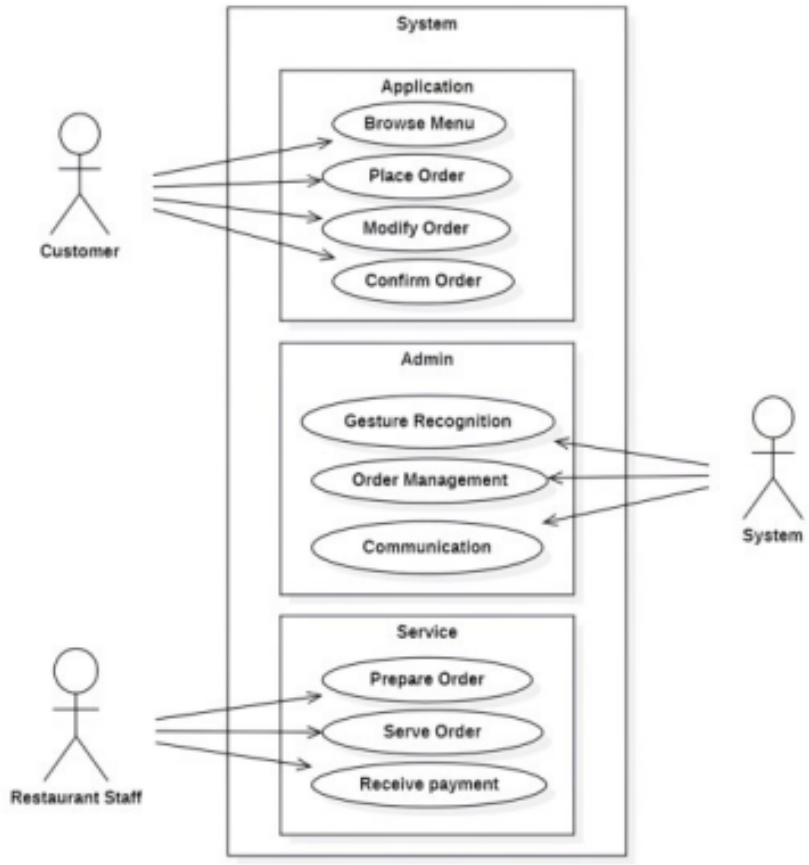
1. Input: extracted features from CNN

2. Tree Construction: construct a tree based on entropy or gini (possibility needs to fine tune it a bit, which may include grid search and prunning)

3. Decision making use the most optimal tree structure to predict hand landmarks for the given data.

2.1.4 System Implementation

After training and testing, the model can be integrated into the system like the diagram.



The research proposed some suggestions on the design of the system to bridge the gesture recognition and actional tasks:

- Interface Layout: a real-time video feed windows; an interactive button to confirm the order; a display pane to show the outcome of the processed gesture.

- System Implementation: gesture capture via a video capture module, process each frame and feed them to the hybrid model for prediction.

- Feedback Display: display the feed back the food predicted in real-time and allow users to correct if there is a mistake in the order and let them rate the predictions.

2.1.5 Quantitative Analysis

For demonstration, they have trained the model on the following labels (gestures). Gestures 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 for Pizza, Breadstick, Burget, Three-layered Sandwich, Four Cheese Pasta, Salad, Six-piece Chicken Nuggets, Seven-layer Dip, Octopus Sushi, Nine-inch Pie, More items. The table below shows us the common metrics use for classification problem

Parameter	***CNN Only***	***Decision Tree***	***Highbred CNN***
Testing Accuracy	87%	80%	91%
Training Accuracy	89%	82%	93%
Precision	90%	83%	92%
Recall	87%	81%	91%
F1-Score	88.5%	82%	91.5%

It can be observed that the hybrid model does better than the other 2 standalone models in all of the metrics. The researchers conclude by saying that although the model shows high accuracy in training, implementing it into restaurants will be a challenge because conditions like lighting, people's handshape, sizes and other unforeseen scenarios need to be taken into considerations.

2.2 Article from Susanto et al. (2022)

The paper proposes a digital restaurant menu system using hand recognition. They divide the research into 6 sections. Section 1 is about the research topic and outline. Section 2 is on different methods to perform gesture recognition and several digital restaurant menus systems. Section 3 is about the research methodology and software development life cycle. Section 4 and 5 is on the design and implementation on the application of gesture recognition module. Section 6 is on the conclusion of the experiment. Section 1 has already been mentioned, so on to section 2.

2.2.1 Method to perform gesture recognition

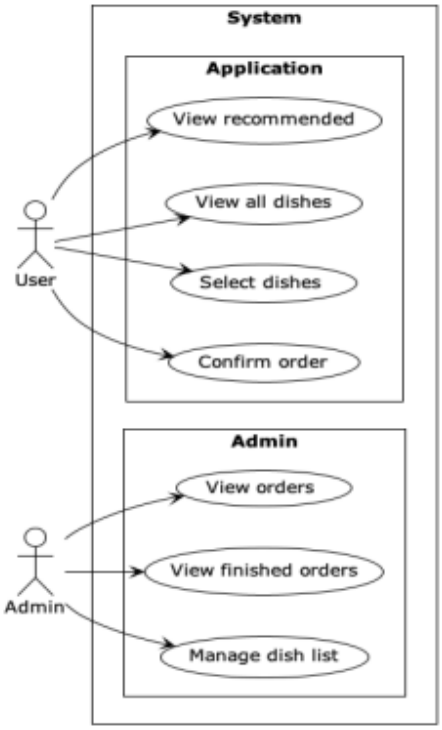
The project used "pose elimination" as the gesture recognition. There are 2 steps in this estimation. The first locate the hand and cropped the image. Next is predict on the image containing only the hand.

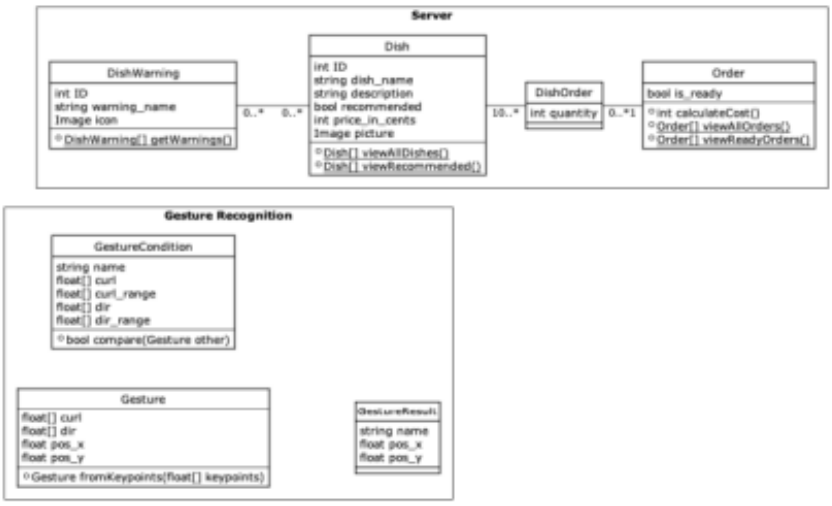
2.2.2 The system development life cycle

The system development life cycle (SDLC) is a process that design, build and deliver to users for them to understand the business's needs. There are 4 main phases in the SDLC, which are planning (plan to design the system), analysis (examines the current system), designing (design the hardware, software, network infrastructure, user interfaces and databases), and finally is the application where the system is built.

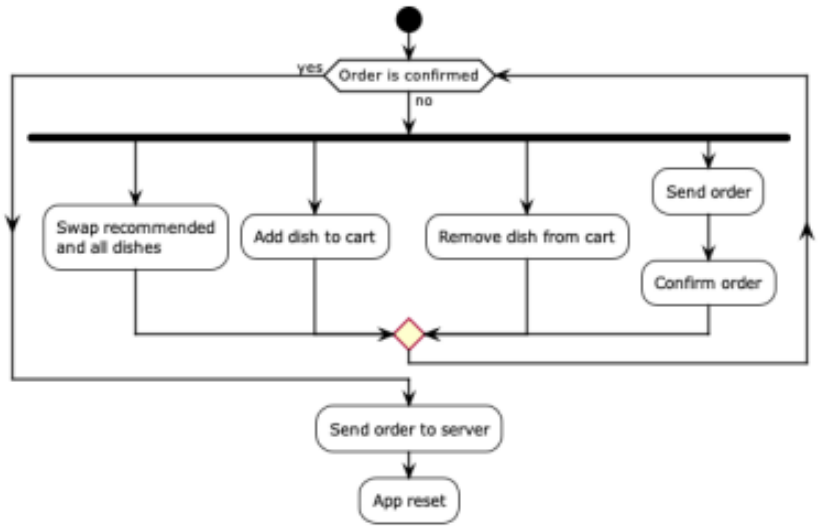
2.2.3 System design

There are 2 main parts which are the customer-oriented and employee-oriented departments. To understand what types of data will be needed, the researchers split the data flow into 2, which are a server that includes all classes used in the digital menu and the Gesture Recognition part includes all classes used in the recognition system.



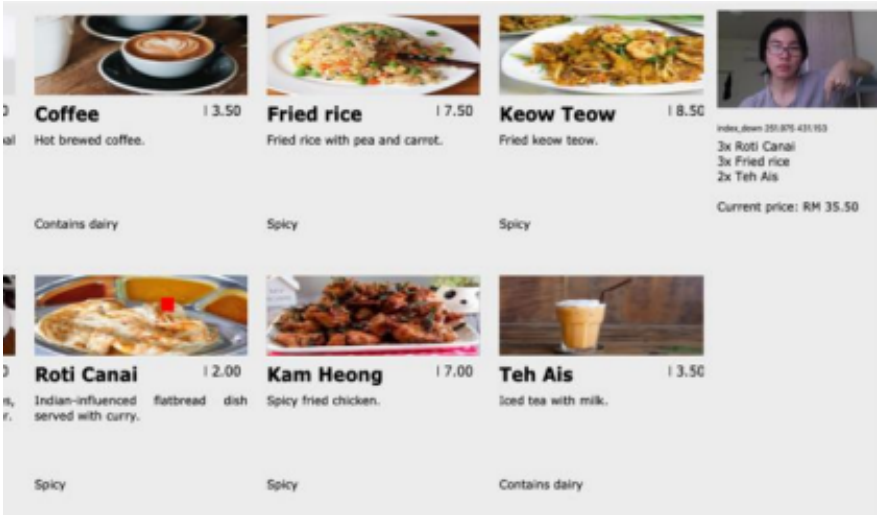


The user's activity is viewed in the diagram below



2.2.4 System Implementation and Evaluation

Implementation



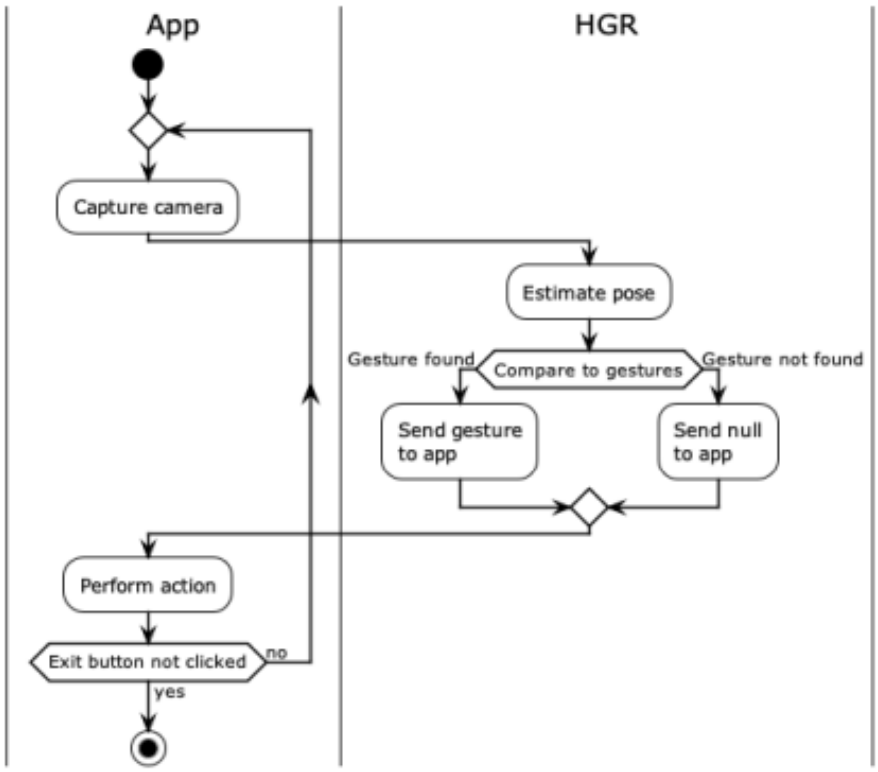
This picture shows the front-end design of the app. Most of the screen is taken by the menu except the top right. There is a camera to view as feedback, below it is the detected gestures and a list of items currently in the cart. The user can interact with the menu as follows. To prevent accidental swapping, any hand sign must be detected for one second. Making a "peace" sign allows user to switch from viewing the recommended dishes and all available dishes. To scroll the menu, the user can hold a "fist" gesture and move left and right. To add an item to the cart, the user performs a "point up" gesture. To add specific item, the user can aim at an image, name or description on the menu. To remove the item, the user can perform "point down". To confirm the order is finished, give a "thumbs up". This will then open a confirmation dialog, which asks the user to either "thumb down" or "thumb up" to confirm or cancel the order. If a customer confirms their order, the number of their order will be shown.

The picture below is the management of orders via web pages.

The left side of the screen shows the order list, and if the order is ready to be served, the staff click "Ready". On the right side, the staff can click "Taken" if the customers have already taken their meal, or "Undo" if the order is canceled. A confirmation message will appear to confirm if the operation is successful or not.

Hand Gesture Recognition

The first thing is to estimate the pose using Mediapipe Hands. Gestures are defined by the curls and direction of each finger. The curl is determined by the average angle between each finger bone, and the direction is measured by the angle between the x-axis and the vector from the base to the tip of the finger. The picture below shows how the application interacts with the hand gesture recognition module.



Evaluation

The researchers propose some evaluations or tests to satisfy the system's requirements. These tests are unit tests, module tests, integration tests, and user acceptance tests. The unit tests can be divided further into user interface, networking module, gesture recognition module, and app control module. The module tests test the interactions between the sub tests of the unit tests. The interaction test is to confirm the interactions between all the modules before integrating everything into the system. Each test is shown on the table in the paper.

2.3 Conclusion

The conclusion shows again the importance of hygienic public places, where not only it is the business' concern, but of customers. Touchless technology can eliminate this problem. They also hope that the implementation of this technology can reduce the maintenance costs and increase efficiency while being hygienic. The researchers have some suggestions on implementing the system. The system consists of the application and server. The application is the all the customer-facing functionalities like viewing menus and ordering. They suggest using the Qt library on C++ for coding the app and Python for building the recognition module. For the server side, Python Django is for the backend, frontend uses HTML, CSS and JavaScript. The overall system is tested against the unit, module and integration tests.

3 My implementation

a) If I plan to build a classification model, I will get the data from various resources like open repos, scrape the internet or create my own dataset.

b) Since machine learning models can only understand numerical data type, I will convert the image into matrices.

c) The data capture protocol can be explained as follows. First, I need to explain the menu label, hand gestures and how to navigate the front-end to place the order.

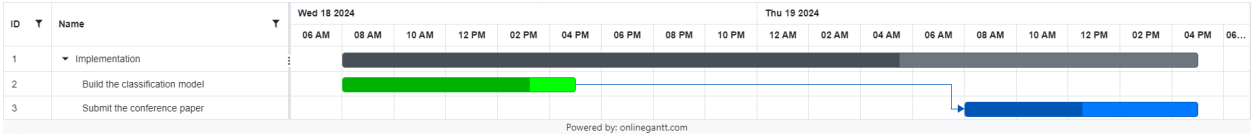
This is the menu, and I have labelled them. The customers can use hand gestures to place the order. The customers would have to hold that gesture for 2 frames for that gesture to be registered. Here is an example of how to order. Gesture "1" to select the beef burger, gesture another "1" to add the Big Mac to the cart. To finish ordering, hold a "thumbs up". The system will ask again to re-confirm the order, hold a "thumbs up" to place order or a "thumbs down" to cancel. The gestures are captured using an Arduino camera module and sent to Python using serial communication. After a frame is captured, it will be loaded into the machine learning model to predict the gestures.

e) The storage type I will use are both cloud and local. For cloud I will use Firebase to save the picture matrices and its label. The same goes for locals. Saving the picture in matrix form will reduce the storage usage because many pictures will take up a lot of space.

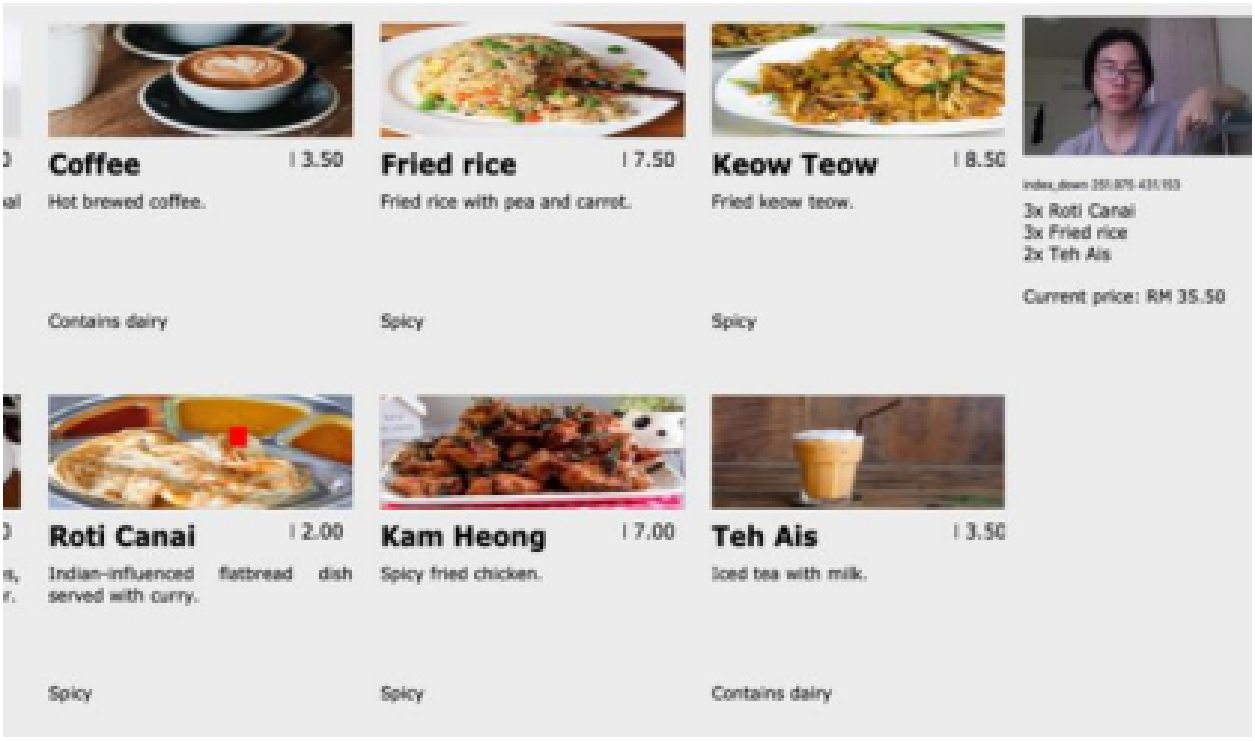
f) For data monitoring, this can be used in the front-end where it will display what the users order live.

g) The data analyses and pattern recognition are part of building the classification model. I plan to use the guide from Shaik et al (2023) to build a hybrid CNN and Decision Tree model. Perhaps I will use Gradient Boost or other ensemble models instead of Decision Tree. The details on how to build the model is explained in literature review section.

i) The hardware I will be using in this case are camera module ov7670 (2.92)and1.8"$tftdisplay$(4.18). The camera module is for capturing, the display is for monitoring the hand gestures in live video. I have already bought them from AliExpress. Here is a rough plan using the Gratt chart.



I can't say for sure, but I will try to build a classification and submit the conference paper before the deadline. The graph said I start on the 18th of this month, and I only have 1 day to finish it. I don't know if I can, but I will try. I can't speak for the whole implementation now, but to finish the whole project, I would need at least 4 days to gather data, build the model, research more on web development to create a simple webpage similar to this.



The front end will display the menu, the customer's live video and their orders. The back end will predict

the gestures and save the data locally and on the cloud. I believe that my current skill set allows me to a project like this since I know a bit about machine learning, but with the tight deadlines as it is right now, I do not know if I can finish this project.

j) There are some potential ethical concerns if this project is implemented in the real world. One is the consent of customers for me to use their data. Data like the customers' hands can be a great source of data to train and improve the model. The second is the barrier to other disabilities. Hand recognition can help the deaf community, but not individuals with other disabilities like those who have motor impairments.

References

Shaik, M., Azam, M., Sindhu, T., Abhilash, K., Mallala, A., & Ganesh, A. (2023). Hand gesture based food ordering system. *2023 International Conference on Smart Systems and Advanced Computing (ICSSAC)*, 867–872. https://doi.org/10.1109/ICSSAS57918.2023.10331637

Susanto, I. C., Subaramaniam, K., & Samad, A. (2022). Restaurant menu with gesture recognition. *Journal of Advances in Artificial Life Robotics*, *3*(2), 102–112. https://doi.org/10.57417/jaalr.3.2_102