**CAN THO UNIVERSITY**

**COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**🙞 🕮 🙜**

**Logo

Description automatically generated**

**GRADUATION THESIS IN**

**INFORMATION TECHNOLOGY**

**(HIGH-QUALITY PROGRAM)**

**EXERCISE CORRECTION WITH MACHINE LEARNING**

**Student: Ngô Hồng Quốc Bảo**

**Student ID: B1809677**

**Class: 2019 - 2023 (K44)**

**Advisor: Prof. Trần Công Án**

**Can Tho, 12/2022**

**CAN THO UNIVERSITY**

**COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**🙞 🕮 🙜**

**Logo

Description automatically generated**

**GRADUATION THESIS IN**

**INFORMATION TECHNOLOGY**

**(HIGH-QUALITY PROGRAM)**

**EXCERCISE CORRECTION WITH MACHINE LEARNING**

**Student: Ngô Hồng Quốc Bảo**

**Student ID: B1809677**

**Class: 2019 - 2023 (K44)**

**Advisor: Prof. Trần Công Án**

**Can Tho, 12/2022**

INSTRUCTOR’S COMMENTS

---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Can Tho, ………………. 2022

(Instructor sign and write full name)

**ACKNOWLEDGMENT**

*I cannot express enough thanks to my instructors/teachers from the College of Information and Communication technology for their continued support and encouragement. My sincere thanks especially go to Prof. Trần Công Án for his guidance and advices throughout the development of this project*.

*My completion of this project could not have been accomplished without the support of my classmates. The supports and advices from them have always been the biggest motivation for me during the whole process. Their encouragement when the times got rough are much appreciated and duly noted.*

*Last but not least, I would like to thank my parents who helped me a lot in gathering different information, collecting data and guiding me, despite their busy schedules, they had always been there with me from time to time in making project.*

Can Tho, December 2022

Student

Ngô Hồng Quốc Bảo

**ABSTRACT**

Fitness is becoming an important part of human life as it brings many beneficial to personal health. However, exercises can also be ineffective and could be dangerous if they are performed incorrectly by the performer. Proper form when doing anything physical is important, but it is especially important when working out or training. Not only can proper form lower your risk of injury but it also allows you to move efficiently, increase your performance, and enables you to have a full range of motion. In my project, I use machine learning to provides detailed analysis and recommendation on the exercises’ performer to improve their form.

In addition, with deep learning and computer vision are being strongly researched and improve by the day. Especially the development of Mediapipe by Google, an open-source framework to “build word-class machine learning solutions” that provides cornerstone Machine Learning models for common tasks like hand tracking, posture detection, ... Make use of the power of Mediapipe’s pose detection, this project “Exercise Correction with Machine Learning” is built in order to analyze, detect and classifying the forms of fitness exercises. The experimental results show that the algorithm proposed in this paper can effectively identify correct and incorrect forms that are performed in an exercise.

**TABLE OF** **CONTENTS**

[INSTRUCTOR’S COMMENTS 1](#_Toc120200254)

[INTRODUCTION 8](#_Toc120200255)

[1. PROBLEM DESCRIPTION 8](#_Toc120200256)

[2. PURPOSE OF THE STUDY 8](#_Toc120200257)

[3. LIMITATION AND SCOPE 9](#_Toc120200258)

[4. GENERAL APPROACH 9](#_Toc120200259)

[5. CRITERIA FOR STUDY SUCCESS 9](#_Toc120200260)

[6. STUDY CONTENTS 9](#_Toc120200261)

[CONTENTS 10](#_Toc120200262)

[CHAPTER 1: AN OVERVIEW OF TECHNOLOGY AND LIBRARY USED 10](#_Toc120200263)

[1. COMPUTER VISION 10](#_Toc120200264)

[1.1. Definition 10](#_Toc120200265)

[1.2. Computer vision common tasks and real-life applications 11](#_Toc120200266)

[2. MEDIAPIPE FRAMEWORK 12](#_Toc120200267)

[2.1. Definition 12](#_Toc120200268)

[2.2. MediaPipe Architecture 13](#_Toc120200269)

[2.3. Advantages of MediaPipe 14](#_Toc120200270)

[3. MEDIAPIPE POSE 15](#_Toc120200271)

[3.1. Overview and Machine Learning pipeline 15](#_Toc120200272)

[3.2. BlazePose detector 16](#_Toc120200273)

[3.3. MediaPipe Pose Output 17](#_Toc120200274)

[4. OTHER TECHNOLOGY AND LIBRARIES USED 17](#_Toc120200275)

[4.1. Open CV 17](#_Toc120200276)

[4.2. Scikit-learn 17](#_Toc120200277)

[4.3. Keras 17](#_Toc120200278)

[4.4. Vue.js 17](#_Toc120200279)

[4.5. Django 17](#_Toc120200280)

[CHAPTER 2: METHODOLOGY 18](#_Toc120200281)

[1. GENERAL APPROACH 18](#_Toc120200282)

[1.1. Exercises selection and errors determination 18](#_Toc120200283)

[1.2. Data Collecting 19](#_Toc120200284)

[1.3. Data processing 20](#_Toc120200285)

[1.4. Model Training 23](#_Toc120200286)

[2. MODEL TRAINING PROCESS IN DEPTH FOR EACH EXERCISES 25](#_Toc120200287)

[2.1. Bicep Curl 25](#_Toc120200288)

[2.2. Basic Plank 26](#_Toc120200289)

[2.3. Basic Squat 28](#_Toc120200290)

[2.4. Lunge 31](#_Toc120200291)

[CHAPTER 3: TESTING AND EXPERIMENTAL RESULTS 34](#_Toc120200292)

[CONCLUSION 35](#_Toc120200293)

[1. Final results 35](#_Toc120200294)

[2. Future works 35](#_Toc120200295)

[REFERENCES 36](#_Toc120200296)

**LIST OF FIGURES**

[Figure 1: Computer Vision example 10](#_Toc120200237)

[Figure 2: Self-driving car using computer vision 12](#_Toc120200238)

[Figure 3: Object detection pipeline in MediaPipe 13](#_Toc120200239)

[Figure 4: Face landmark detection and segmentation using MediaPipe 14](#_Toc120200240)

[Figure 5: The 33 landmarks model in MediaPipe Pose predicts 15](#_Toc120200241)

[Figure 6: MediaPipe pose detection's pipeline 16](#_Toc120200242)

[Figure 7: Example of pose detection by MediaPipe 17](#_Toc120200243)

[Figure 8: A person doing lunge 18](#_Toc120200244)

[Figure 9: Example of a directory tree structure for each exercise collected videos 20](#_Toc120200245)

[Figure 10: Example of important landmarks for plank exercise 21](#_Toc120200246)

[Figure 11: Class balance of Bicep Curl's dataset 26](#_Toc120200247)

[Figure 12: Class balance of Basic Plank's dataset 27](#_Toc120200248)

[Figure 13: Class balance of Squat's dataset 29](#_Toc120200249)

[Figure 14: Graph representing the ratio between feet width and shoulder width in squat 30](#_Toc120200250)

[Figure 15: Graph representing the ratio between knee width and feet width in each stage of squat 31](#_Toc120200251)

[Figure 16: Graph representing the angles of right and left knee in correct lunge position 32](#_Toc120200252)

[Figure 17: Class balance of Lunge’s dataset 33](#_Toc120200253)

**LIST OF TABLES**

No table of figures entries found.

INTRODUCTION

1. PROBLEM DESCRIPTION

The advancements in artificial intelligence and computational power, computer vision technology has taken a huge leap toward integration in our daily lives. Computer vision is the field of computer science that focuses on creating digital systems that can process, analyze, and make sense of visual data (images or videos) in the same way that humans do. The concept of computer vision is based on teaching computers to process an image at a pixel level and understand it. Technically, machines attempt to retrieve visual information, handle it, and interpret results through special software algorithms.

In addition, nowadays, exercise is proven to be extremely important for people’s daily lives. Being physically active can improve your brain health, help manage weight, reduce the risk of disease, strengthen bones and muscles, and improve your ability to do everyday activities. Everyone can experience the health benefits of physical activity regardless of their age, gender, ethnicity, shape ... When it comes to exercises quality is more important than quantity. How an exercise is performed may mean the difference between going harder and getting sidelined. Anyone the more experienced can benefit from some occasional form feedback. Perfecting form will boost performance, conserve energy and reduce injuries over time.

However, for many people, especially beginners or people who mainly do their exercise by themselves, a noticeable amount of them perform these exercises improper form due to the lack of awareness or instruction. With improper form, you risk straining/injuring your body rather than training it. One thing is for sure, you are not going to see results if you are sidelined with a frustrating injury.

This thesis is intended to build machine learning models for different exercises with the help of Mediapipe framework. For each exercise, the correspond model will detect whether if the exercise is performed properly or not. The trained model will be applied to build a web application where users can provide their workout videos to get feedbacks on their form.

1. PURPOSE OF THE STUDY

The purpose of this study is to develop 4 machine learning models for 4 of the most home exercises which each model can detect any form of incorrect movement while a person is performing a correspond exercise. In addition, a web application that utilize the trained models, will be built in other to analyze and provide feedbacks on workout videos.

1. LIMITATION AND SCOPE

This study researches deep learning knowledge such as computer vision, neural network and Mediapipe framework. The study also uses Python programming language, Open CV library for image processing; Sci-kit learn library and Keras library for building machine learning model; Vue.js framework and Django framework for the web application that utilizes the trained model for feedbacks on incorrect form to exercise videos.

1. GENERAL APPROACH

* Research on which popular exercises which commonly improperly perform.
* Research on which technology to choose that is suitable to solve the problem.
* Collect and process data of the chosen exercises.
* Train and evaluate model for each exercise.
* Build a web application utilize all the trained models.
* Evaluate the final model and test the web application.

1. CRITERIA FOR STUDY SUCCESS

* Successfully build 4 models for 4 exercises which can detect proper and improper form of each exercise.
* Successfully build a web application which apply the trained models for providing feedbacks and recommendation for users bases on their exercise videos.
* Earn knowledge on deep learning and computer vision topics. Apply Mediapipe framework to real life application.

1. STUDY CONTENTS

- **Introduction**: An overview of the thesis: an introduction to the topic, research methods and layout of the thesis.

- **Content part**: The content of the thesis is divided into 3 chapters

+ Chapter 1: Required specifications

+ Chapter 2: An overview of technology used

+ Chapter 3: System descriptions

+ Chapter 4: Delivery system design

+ Chapter 5: Testing and experimental results

- **Conclusion**: Present the results achieved and the development direction of the system

CONTENTS

1. AN OVERVIEW OF TECHNOLOGY AND LIBRARY USED

1. COMPUTER VISION

1.1. Definition

Computer vision is the field of computer science that focuses on creating digital systems that can process, analyze, and make sense of visual data (images or videos) in the same way that humans do. The concept of computer vision is based on teaching computers to process an image at a pixel level and understand it. Technically, machines attempt to retrieve visual information, handle it, and interpret results through special software algorithms.

Computer vision trains machines to perform these functions, but it has to do it in much less time with cameras, data and algorithms rather than retinas, optic nerves and a visual cortex. Because a system trained to inspect products or watch a production asset can analyze thousands of products or processes a minute, noticing imperceptible defects or issues, it can quickly surpass human capabilities.

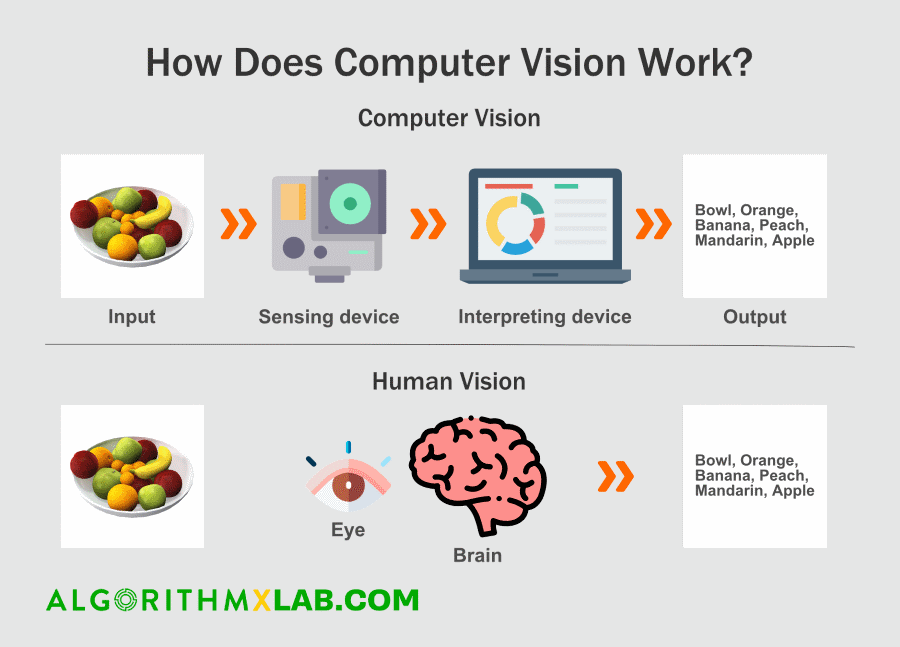


Figure 1: Computer Vision example

Computer vision needs lots of data. It runs analyses of data over and over until it discerns distinctions and ultimately recognize images. For example, to train a computer to recognize automobile tires, it needs to be fed vast quantities of tire images and tire-related items to learn the differences and recognize a tire, especially one with no defects.

1.2. Computer vision common tasks and real-life applications

In essence, computer vision tasks are about making computers understand digital images as well as visual data from the real world. This can involve extracting, processing, and analyzing information from such inputs to make decisions. The most popular computer vision tasks that we regularly find in AI jargon include:

* *Image Classification*: Given a group of images, the task is to classify them into a set of predefined classes using solely a set of sample images that have already been classified.
* *Object detection*: refers to detection and localization of objects using bounding boxes. Object detection looks for class-specific details in an image or a video and identifies them whenever they appear.
* *Image segmentation*: the division of an image into subparts or sub-objects to demonstrate that the machine can discern an object from the background and/or another object in the same image.
* *Face and person recognition*: where features are detected and localized, facial recognition performs not only detection, but also recognition of the detected face.
* *Video motion analysis*: refers to the study of moving objects or animals and the trajectory of their bodies.

There is a lot of research being done in the computer vision field, but it’s not just research. Real-world applications demonstrate how important computer vision is to endeavors in business, entertainment, transportation, healthcare and everyday life. A key driver for the growth of these applications is the flood of visual information flowing from smartphones, security systems, traffic cameras and other visually instrumented devices.

* Google Translate lets users point a smartphone camera at a sign in another language.
* IBM is applying computer vision technology with partners like Verizon to bring intelligent AI to the edge, and to help automotive manufacturers identify quality defects before a vehicle leaves the factory.
* The development of self-driving vehicles relies on computer vision to make sense of the visual input from a car’s cameras and other sensors. It’s essential to identify other cars, traffic signs, lane markers, pedestrians, bicycles and all of the other visual information encountered on the road.



Figure 2: Self-driving car using computer vision

2. MEDIAPIPE FRAMEWORK

2.1. Definition

MediaPipe is an open-source framework for building pipelines to perform computer vision inference over arbitrary sensory data such as video or audio. Using MediaPipe, such a perception pipeline can be built as a graph of modular components. MediaPipe was built for machine learning (ML) teams and software developers who implement production-ready ML applications, or students and researchers who publish code and prototypes as part of their research work.

In computer vision pipelines, those components include model inference, media processing algorithms, data transformations, etc. Sensory data such as video streams enter the graph, and perceived descriptions such as object-localization or face-keypoint streams exit the graph.

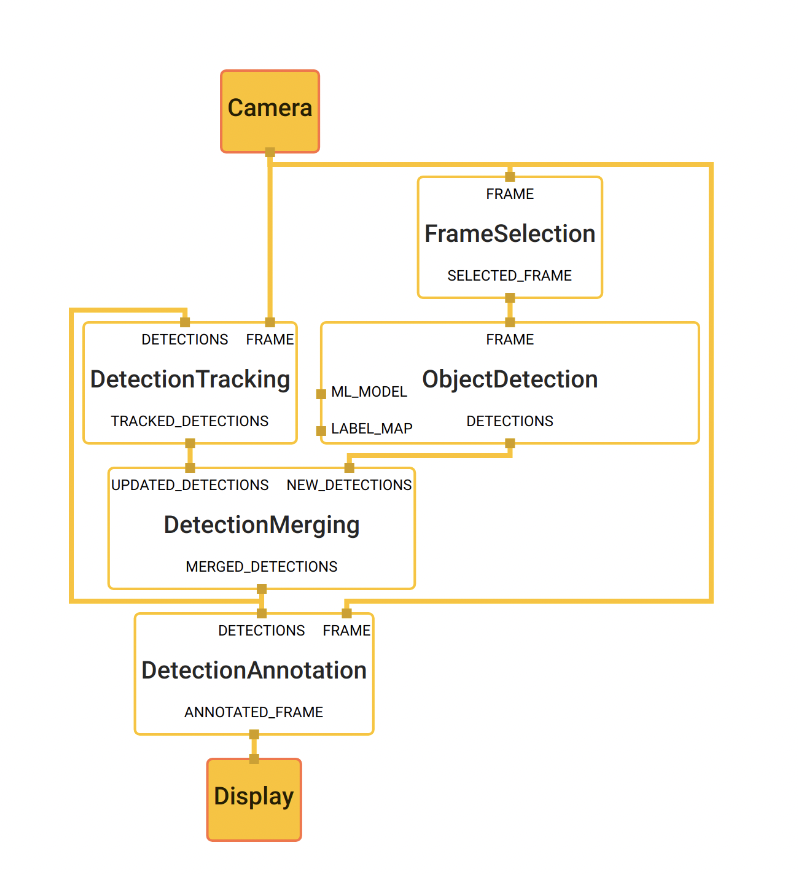


Figure 3: Object detection pipeline in MediaPipe

The main use case for MediaPipe is rapid prototyping of perception pipelines with inference models and other reusable components. MediaPipe also facilitates the deployment of perception technology into demos and applications on a wide variety of different hardware platforms. MediaPipe enables incremental improvements to perception pipelines through its rich configuration language and evaluation tools.

2.2. MediaPipe Architecture

MediaPipe allows a developer to prototype a pipeline incrementally. A pipeline is defined directed graph of components where each component is a *Calculator*. The graph is specified using a *GraphConfig* *protocol* buffer and then run using a Graph object.

In the graph, the calculators are by data, *Streams*. Each represents a time-series of data, *Packets*. Together, the calculators and streams define a data-flow graph. The packets which flow across the graph are collated by their timestamps within the time-series.

* *Packet*: basic data unit in MediaPipe. A packet consists of a numeric timestamp and a shared pointer to an immutable payload.
* *Stream*: each node in the graph is connected to another node through a stream. A stream carries a sequence of packet whose timestamps must be monotonically increasing.
* *Calculators*: implemented as each node of a graph. The bulk of graph execution happens inside its calculator.
* *Graph*: all processing takes places within the context of a Graph. A graph contains a collection of nodes joined by directed connections along which packets can flow.
* *GraphConfig*: is a specification that describes the topology and functionality of a MediaPipe graph. All the necessary configurations of the node, such its type, inputs and outputs must be described in the specification. Description of the node can also include several optional fields, such as node-specific options, input policy and executor.

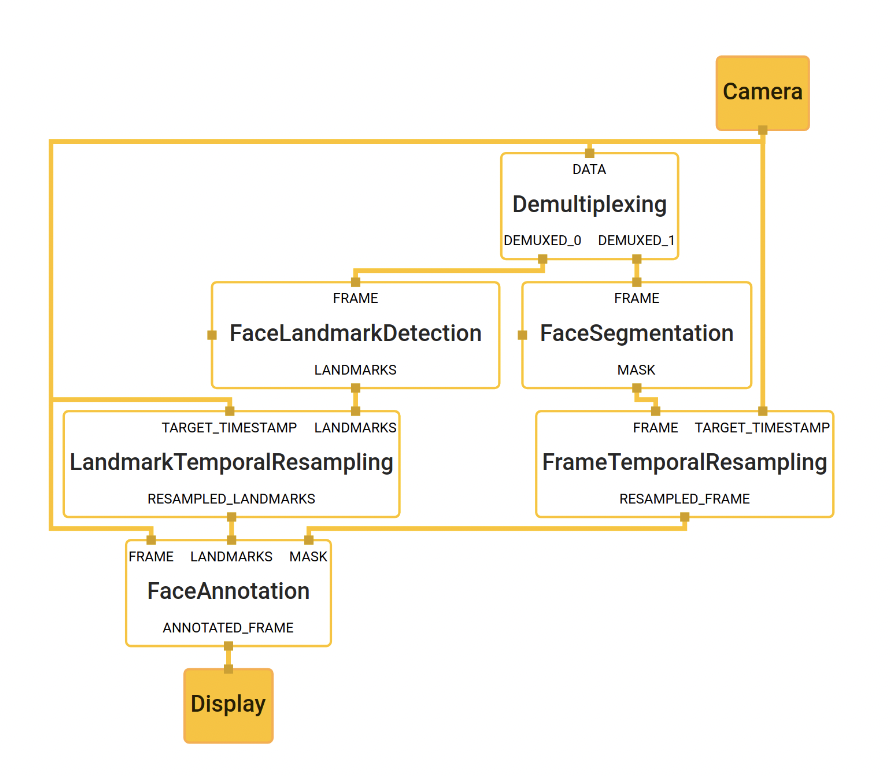


Figure 4: Face landmark detection and segmentation using MediaPipe

2.3. Advantages of MediaPipe

* *End-to-end acceleration*: Use common hardware to build-in fast ML inference and video processing, including GPU, CPU, or TPU.
* *Build once, deploy anywhere*: The unified framework is suitable for Android, iOS, desktop, cloud, web, and IoT platforms.
* *Ready-to-use solutions*: Prebuilt ML solutions demonstrate the full power of the MediaPipe framework.
* *Open source and free*: The framework is licensed under Apache 2.0, fully extensible, and customizable.

3. MEDIAPIPE POSE

3.1. Overview and Machine Learning pipeline

MediaPipe Pose is a ML solution for high-fidelity body pose tracking, inferring 33 3D landmarks and background segmentation mask on the whole body from RGB video frames utilizing our BlazePose research that also powers the ML Kit Pose Detection API. Current state-of-the-art approaches rely primarily on powerful desktop environments for inference, whereas our method achieves real-time performance on most modern mobile phones, desktops/laptops, in python and even on the web.



Figure 5: The 33 landmarks model in MediaPipe Pose predicts

The solution utilizes a two-step detector-tracker ML pipeline, proven to be effective in MediaPipe Hands and MediaPipe Face Mesh solutions. Using a detector, the pipeline first locates the person/pose region-of-interest (ROI) within the frame. The tracker subsequently predicts the pose landmarks and segmentation mask within the ROI using the ROI-cropped frame as input. Note that for video use cases the detector is invoked only as needed, i.e., for the very first frame and when the tracker could no longer identify body pose presence in the previous frame. For other frames the pipeline simply derives the ROI from the previous frame’s pose landmarks.

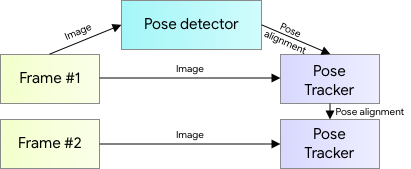


Figure 6: MediaPipe pose detection's pipeline

3.2. BlazePose detector

For real-time performance of the full ML pipeline consisting of pose detection and tracking models, each component must be very fast, using only a few milliseconds per frame. To accomplish this, it is known that the strongest signal to the neural network about the position of the torso is the person's face (due to its high-contrast features and comparably small variations in appearance). Therefore, BlazePose is a fast and lightweight pose detector by making the strong (yet for many mobile and web applications valid) assumption that the head should be visible for single-person use case.

Therefore, the detector is inspired by the lightweight BlazeFace model, used in MediaPipe Face Detection, as a proxy for a person detector. It explicitly predicts two additional virtual keypoints that firmly describe the human body center, rotation and scale as a circle. Inspired by Leonardo’s Vitruvian man, the model predicts the midpoint of a person’s hips, the radius of a circle circumscribing the whole person, and the incline angle of the line connecting the shoulder and hip midpoints.

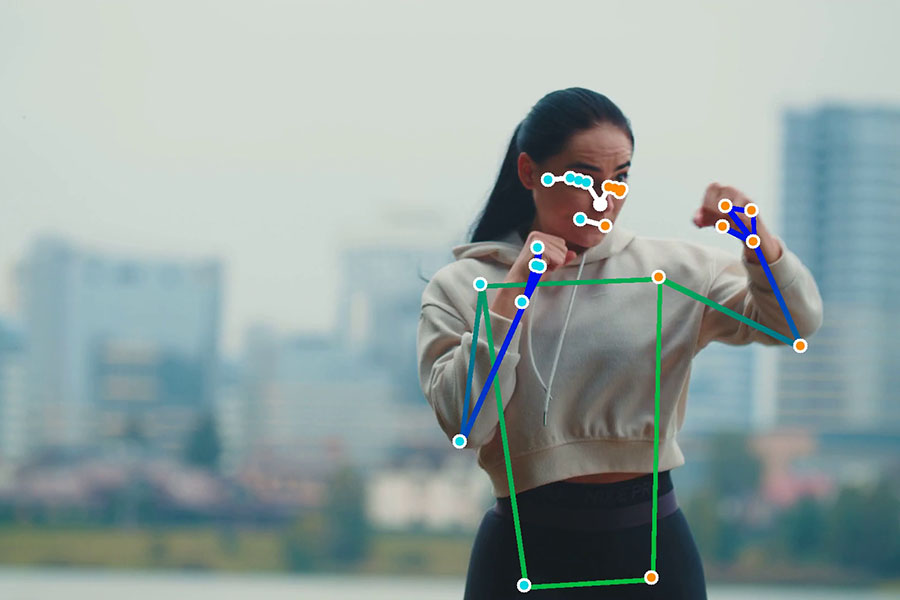


Figure 7: Example of pose detection by MediaPipe

3.3. MediaPipe Pose Output

MediaPipe Pose returns a list of 33 pose landmarks. Each landmark consists of the following properties:

* *x* and *y*: Landmark coordinates normalized to [0.0, 1.0] by the image width and height respectively.
* *z:* Represents the landmark depth with the depth at the midpoint of hips being the origin, and the smaller the value the closer the landmark is to the camera. The magnitude of *z* uses roughly the same scale as *x*.
* *visibility*: A value in *[0.0, 1.0]* indicating the likelihood of the landmark being visible (present and not occluded) in the image.

4. OTHER TECHNOLOGY AND LIBRARIES USED

4.1. Open CV

4.2. Scikit-learn

4.3. Keras

4.4. Vue.js

4.5. Django

1. METHODOLOGY

1. GENERAL APPROACH

For every exercise, the following steps will be applied for its error detections:

* Error determination.
* Data collecting (videos, public dataset).
* Converting collected data to a .csv file.
* Train a model/models for error detection.

1.1. Exercises selection and errors determination

First and foremost, the initial step of this thesis is to decide on which exercises to train machine learning model for form correction. Due to the limitation of time for the thesis, only 4 exercises are picked. Each exercise needs to meet the following criteria:

* An exercise must be a common exercise.
* An exercise is popular among people who workouts at home or workouts mostly by themselves. Therefore, it would raise the possibility that the exercise would be performed improperly.
* An exercise must contain at least 2 popular errors which impact the inefficiency of the exercise.

Depends on the above criteria, I decided on 4 exercises: *bicep curl*, *basic plank*, *basic quat* and *lunge*.



Figure 8: A person doing lunge

After decided on 4 exercises, for each exercise, at least 2 popular errors for that exercise would be determined and then develop strategy to detect those errors (The detail of each error will be going in depth in this section). Here are the determined common errors for each exercise:

* *Bicep Curl*: Loose upper arm, weak peak contraction and lean-back standing posture.
* *Basic plank*: Lower back and High back.
* *Basic squat*: Foot placement too tight/wide and knee placement too wide/tight.
* *Lunge*: Knee over toe while going down.

1.2. Data Collecting

To be able to build exercise correction machine learnings model for this thesis, for each exercise, there are steps to be followed: collecting data (online videos, self-made video, online dataset), processing data depends on the exercise, training and evaluating model.

1.2.1. Self-collected data

Due to the lack of videos or dataset online that recorded human doing exercises both in a proper or improper way, the majority of self-collected videos were either recorded by myself, my friends or my family. In total, there are 5 contributors for this dataset. For each exercise, each contributor is required to:

* Record 2 videos with different camera angles when the contributor performs the exercise correctly.
* Record 2 videos with different camera angles for each determined error of the exercises.

For each video that provided by each contributor, the video has to meet these following criteria:

* Minimum duration: depends on exercise. With exercises which does not require much action such as plank, the minimum required duration of a video would be 30 seconds. However, with exercises which involve more movements such as squat, the contributors are required to perform the exercise at least 15 reps for each video. (More detail in …)
* Video is recorded in environment with enough light so that the whole body of a person is in good visibility.
* For each exercise, video needs to show clearly the important joints or movement which important to the exercise.
* The contributor needs to be on frame throughout the video.

After collecting the videos, depends on the exercise, they are divided into multiple classes:

* “proper-form”: Videos of a contributor perform the exercise correctly.
* “error-name”: Each video of a contributor performs the exercise correctly will be in a separate class.

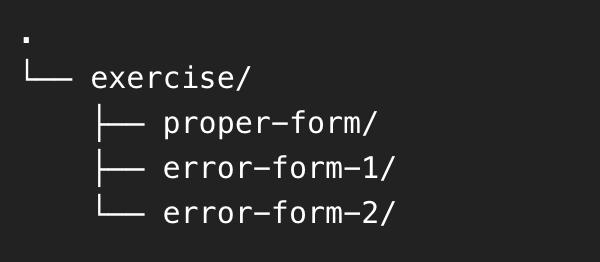


Figure 9: Example of a directory tree structure for each exercise collected videos

1.2.2. Public Dataset for Kaggle

With an exercise such as *Plank*, as there is not much movement during the exercise, I’m able to find a dataset from an open database from Kaggle. In short, Kaggle is an online community platform for data scientists and machine learning enthusiasts.

The found dataset is about many yoga poses but the very well-known ones are the downward dog pose, goddess pose, tree pose, plank pose and the warrior pose. The dataset contains 5 folders for 5 poses, each folder contains images of people correctly doing the correspond pose.

For the purpose of this thesis, only the folder contains the images of people properly doing plank is chosen. There are 266 image files in that folder, I handpicked all the images that represent a basic plank and discard the reset. In conclusion, there are 30 images which are arranged to the proper form class for *basic plank*.

1.3. Data processing

In general, the operation of processing the self-collected video is to identify human pose with MediaPipe Pose and then store data in an *.csv* file. The operation is particular depends on which exercise, however, there are a few basic steps which are handled the same throughout all 4 exercises, this section will discuss about these steps.

1.3.1. Detecting important landmarks

For each exercise, there will be different poses/body’s position, therefore it is essential to identify which parts (shoulder, hip, …) of a body are contribute to the exercise. The important landmarks identified for each exercise are utilized to extract body part’s position while exercising using MediaPipe.

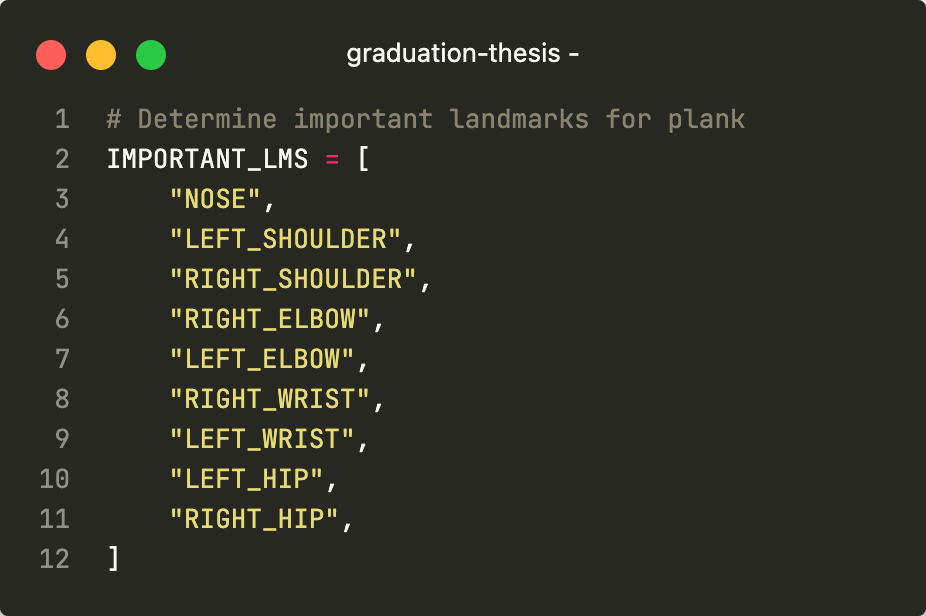


Figure 10: Example of important landmarks for plank exercise

In addition, every landmark’s property such as *x, y, z* and *visibility*, will be flatten to 4 headers for an *csv* file. For example, with the *NOSE* landmarks, the 4 headers would be: NOSE\_x, NOSE\_y, NOSE\_z and NOSE\_v(detail mentioned in page 17). Therefore, for each exercise, an empty *csv* file will be initialized. The first header is the “label” which contain a class for each datapoint, the rest of file’s headers are all the important headers with their properties flatten.

label,nose\_x,nose\_y,nose\_z,nose\_v

C,0.549,0.32,-0.175,0.992

L,0.538,0.301,-0.178,0.994

1.3.2. Extract data from video using OpenCV and MediaPipe

With every exercise, there are videos which are separated into different classes. With each video, every frame that matches a certain form of an exercise would be:

* Pose estimation: Process the video using OpenCV then utilize MediaPipe Pose to produce a list of coordinate predictions of all keypoint locations, and their corresponding prediction confidence.

Configuration options for Medipipe Pose:

* + STATIC\_IMAGE\_MODE: If set to false, the solution treats the input images as a video stream. It will try to detect the most prominent person in the very first images, and upon a successful detection further localizes the pose landmarks. In subsequent images, it then simply tracks those landmarks without invoking another detection until it loses track, on reducing computation and latency. If set to true, person detection runs every input image, ideal for processing a batch of static, possibly unrelated, images. Default to false.
  + MODEL\_COMPLEXITY: Complexity of the pose landmark model: 0, 1 or 2. Landmark accuracy as well as inference latency generally go up with the model complexity. Default to 1.
  + SMOOTH\_LANDMARKS: If set to true, the solution filters pose landmarks across different input images to reduce jitter, but ignored if static\_image\_mode is also set to true. Default to true.
  + ENABLE\_SEGMENTATION: If set to true, in addition to the pose landmarks the solution also generates the segmentation mask. Default to false.
  + SMOOTH\_SEGMENTATION: If set to true, the solution filters segmentation masks across different input images to reduce jitter. Ignored if enable\_segmentation is false or static\_image\_mode is true. Default to true.
  + MIN\_DETECTION\_CONFIDENCE: Minimum confidence value ([0.0, 1.0]) from the person-detection model for the detection to be considered successful. Default to 0.5.
  + MIN\_TRACKING\_CONFIDENCE: Minimum confidence value ([0.0, 1.0]) from the landmark-tracking model for the pose landmarks to be considered tracked successfully, or otherwise person detection will be invoked automatically on the next input image. Setting it to a higher value can increase robustness of the solution, at the expense of a higher latency. Ignored if static\_image\_mode is true, where person detection simply runs on every image. Default to 0.5.

cap = cv2.VideoCapture("/path/to/video")

with mp\_pose.Pose(min\_detection\_confidence=0.5, min\_tracking\_confidence=0.5) as pose:

while cap.isOpened():

ret, image = cap.read()

image = cv2.cvtColor(image, cv2.COLOR\_BGR2RGB)

results = pose.process(image)

* For every frame with the list of predicted landmarks, the important landmarks for a correspond exercise would be extracted in append as a row to the csv file with the label column matches the class of the video.
* The collected data which is saved in a csv file would be split to train and evaluate model. 80% of the data will be used for model training and the remaining 20% will be used for model evaluation.

1.4. Model Training

There are 2 methods used in this thesis for model training. For each exercise, the models trained for each method will be compared and the best model will be chosen.

* Classification with Scikit-learn.
* Building a Neural Network for classification with Keras.

1.4.1. Classification with Scikit-learn

The machine learning algorithms used for model training are: Decision Tree/Random Forest, K-Nearest Neighbors (KNN), C-Support Vector (SVC), Linear classifiers (SVM, logistic regression, etc.) with SGD training and Classifier using Ridge regression.

After the training session, each result from each algorithm will be evaluate with metrics such as: precision score, accuracy score and F1 score. The algorithm which provides the best results will be selected. Below is the sample code for this section.

algorithms = [("LR", LogisticRegression()),

("SVC", SVC()),

('KNN',KNeighborsClassifier()),

("DTC", DecisionTreeClassifier()),

("SGDC", SGDClassifier()),

("Ridge", RidgeClassifier()),

('RF', RandomForestClassifier()),]

models = {}

for name, model in algorithms:

trained\_model = model.fit(X\_train, y\_train)

models[name] = trained\_model

model\_results = model.predict(X\_test)

p\_score = precision\_score(y\_test, model\_results,

average="macro")

a\_score = accuracy\_score(y\_test, model\_results)

f1\_score\_result = f1\_score(y\_test, model\_results,

average=None, labels=["C", "L", "H"])

1.4.2. Neural Network for classification with Keras

With Keras, composing a neural network by creating layers and linking them together. In this project, one type of layer called a fully connected or Dense layer. In Keras, this is defined by the keras.layers.Dense class.

A dense layer has a number of neurons, which is a parameter which can be chosen when the layer is created. When connecting the layer to its input and output layers every neuron in the dense layer gets an edge (i.e. connection) to all of the input neurons and all of the output neurons.

Occasionally, I used another type of layer which is Dropout layer. The term “dropout” refers to dropping out the nodes (input and hidden layer) in a neural network. The nodes are dropped by a dropout probability of p. Dropout layer is used to combat the problem of overfitting. In overfitting, a unit may change in a way that fixes up the mistakes of the other units, using dropout, it prevents these units to fix up the mistake of other units, thus preventing co-adaptation, as in every iteration the presence of a unit is highly unreliable. Therefore, by randomly dropping a few units (nodes), it forces the layers to take more or less responsibility for the input by taking a probabilistic approach.

One of a parameter for a hidden layer is activation function. In short, activation function in the hidden layer will control how well the network model learns the training dataset. The choice of activation function in the output layer will define the type of predictions the model can make. In this thesis, there are only 2 activation functions which is used:

* relu function: Applies the rectified linear unit activation function. With default values, this returns the standard ReLU activation: max(x, 0), the element-wise maximum of 0 and the input tensor..
* softmax function: converts a vector of values to a probability distribution. The elements of the output vector are in range (0, 1) and sum to 1.

Below is a sample code for a neural network built with Keras.

model = Sequential()

model.add(Dense(68, input\_dim = 68, activation = "relu")) model.add(Dropout(0.5))

model.add(Dense(68, activation = "relu")) model.add(Dropout(0.5))

model.add(Dense(14, activation = "relu"))

model.add(Dense(3, activation = "softmax")) model.compile(Adam(lr = 0.01), "categorical\_crossentropy", metrics = ["accuracy"])

model.summary()

2. MODEL TRAINING PROCESS IN DEPTH FOR EACH EXERCISES

The previous section discusses about the general approach, each step to build models for the 4 exercises, this section will address in detail on each exercise. neurons and all of the output neurons.

2.1. Bicep Curl

2.1.1. Basic technique, errors description and important landmarks

The biceps curl is a highly recognizable weight-training exercise that works the muscles of the upper arm, and to a lesser extent, those of the lower arm. The exercise has to be performed while standing, involve around the arms moving up and down while holding a dumbbell or barbell.

There are 3 popular errors of bicep curl that will be targeted in this thesis:

* Loose upper arm: when an arm moves upward during the exercise, the upper arm is moving instead of staying still.
* Weak peak contraction: when an arm moves upward, it does not go high enough therefore not put enough contraction to the bicep.
* Rocking standing posture: the performer’s torso leans back and fore during the exercise for momentum.

In my research and exploration, the important MediaPipe Pose landmarks for this exercise are: nose, left shoulder, right shoulder, right elbow, left elbow, right wrist, left wrist, right hip and left hip.

2.1.2. Errors detection method

a. Loose upper arm

Can be detected by calculating the angle between the elbow, shoulder and the shoulder’s projection on the ground. Through my research, if the angle is over 40 degrees, the movement will be classified as a “loose upper arm” error.

b. Weak peak contraction

Can be detected by calculating the angle between the wrist, elbow and shoulder when the performer’s arm is coming up. Through my research, if the angle is more than 60 degrees before the arm comes down, the movement will be classified as a “weak peak contraction” error.

c. Rocking standing posture

Due to its complexity, machine learning is used for this error. Videos of contributors perform with and without this error are used as data for training model. Figure 11 is a visual graph represent the number of frames gathered from the videos and their classes. There are a total of 15372 images, in which, there are 8238 samples (53.6%) belong to class *correct posture (C)* and 7134 (46.4%) samples belong to class *rocking posture (L)*.

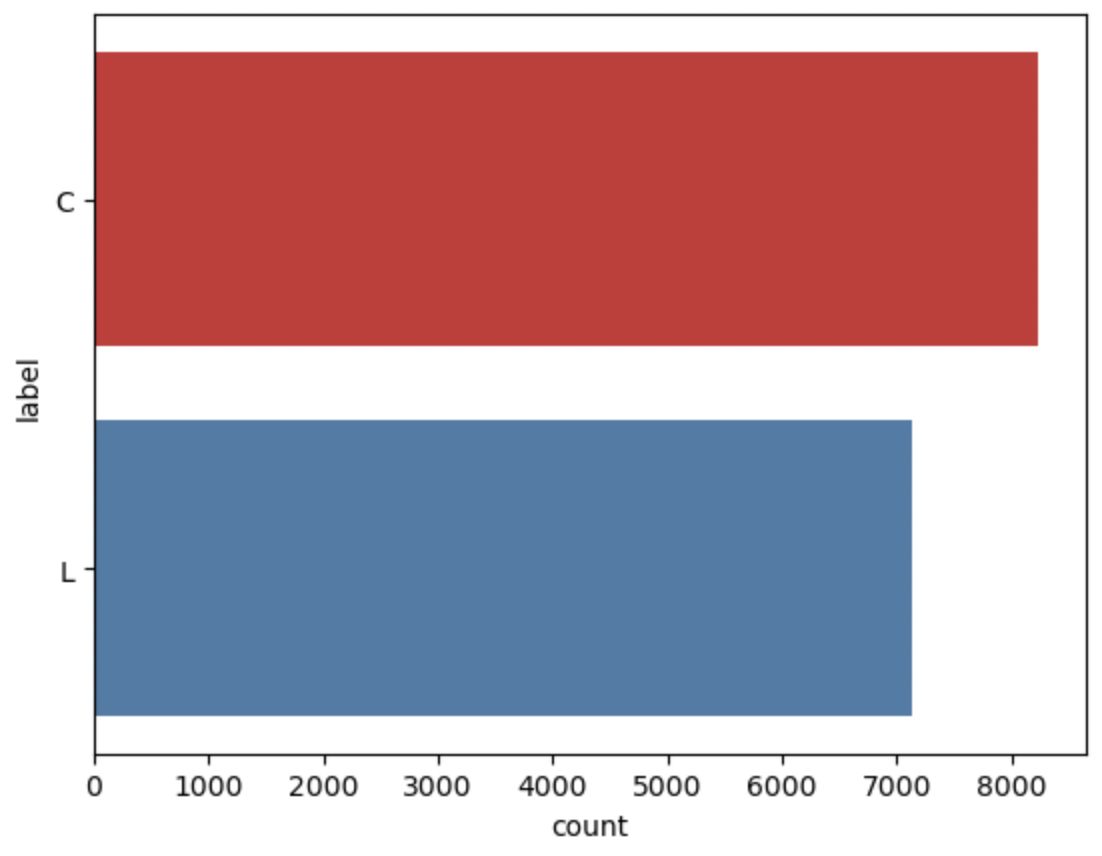


Figure 11: Class balance of Bicep Curl's dataset

Every datapoint in the dataset is consist of 37 columns which are formed based on its classed and the important landmarks. Below is an example of a datapoint.

label,nose\_x,nose\_y,nose\_z,nose\_v,left\_shoulder\_x,left\_shoulder\_y,left\_shoulder\_z,left\_shoulder\_v,right\_shoulder\_x,right\_shoulder\_y,right\_shoulder\_z,right\_shoulder\_v,right\_elbow\_x,right\_elbow\_y,right\_elbow\_z,right\_elbow\_v,left\_elbow\_x,left\_elbow\_y,left\_elbow\_z,left\_elbow\_v,right\_wrist\_x,right\_wrist\_y,right\_wrist\_z,right\_wrist\_v,left\_wrist\_x,left\_wrist\_y,left\_wrist\_z,left\_wrist\_v,left\_hip\_x,left\_hip\_y,left\_hip\_z,left\_hip\_v,right\_hip\_x,right\_hip\_y,right\_hip\_z,right\_hip\_v

C,0,0.55,0.32,-0.18,0.99,0.49,0.34,0.13,0.99,0.49,0.34,-0.24,1.0,0.47,0.51,-0.27,0.96,0.46,0.5,0.15,0.02,0.48,0.68,-0.22,0.83,0.46,0.63,0.06,0.03,0.43,0.64,0.12,0.95,0.44,0.65,-0.12,0.97

2.2. Basic Plank

2.2.1. Basic technique, errors description and important landmarks

The plank, or planking, is an exercise that involves your core muscles, improving your strength, balance and endurance. To perform a plank, lying on the ground with the elbows in line with the shoulder and the feet shoulder width apart, Push body up bearing the weight on the forearms and feet, body is kept straight.

There are 3 popular errors of basic plank that will be targeted in this thesis:

* High lower back: while performing the exercise, instead of keeping the lower back straight, it is raised too high.
* Low lower back: while performing the exercise, instead of keeping the lower back straight, it is brought down too low.

In my research and exploration, the important MediaPipe Pose landmarks for this exercise are: nose, left shoulder, right shoulder, right elbow, left elbow, right wrist, left wrist, right hip, left hip, right knee, left knee, right ankle, left ankle, right heel, left heel, right foot index and left foot index.

2.2.2. Errors detection method

Videos of contributors perform in both 3 stages (proper form, high lower back and low lower back) of this exercise are used for to build a machine learning model. Figure 12 is a visual graph represent the number of frames gathered from the videos and their classes. There are a total of 28623 images, in which, there are 9630 samples (33.6%) belong to class *correct form (C)*, 8982 samples (31.4%) belong to class *high back (H)* and 10011 (35%) samples belong to class *low back (L).*

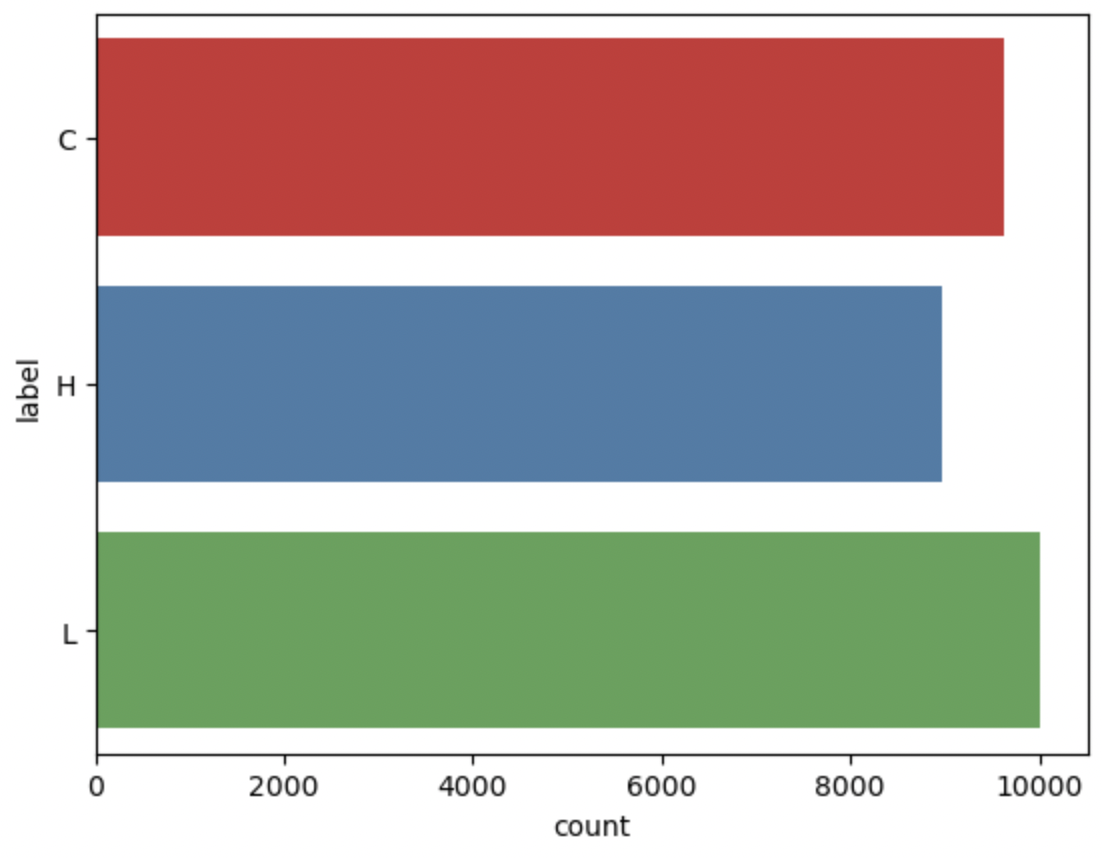


Figure 12: Class balance of Basic Plank's dataset

Every datapoint in the dataset is consist of 69 columns which are formed based on its classed and the important landmarks. Below is an example of a datapoint:

label,nose\_x,nose\_y,nose\_z,nose\_v,left\_shoulder\_x,left\_shoulder\_y,left\_shoulder\_z,left\_shoulder\_v,right\_shoulder\_x,right\_shoulder\_y,right\_shoulder\_z,right\_shoulder\_v,left\_elbow\_x,left\_elbow\_y,left\_elbow\_z,left\_elbow\_v,right\_elbow\_x,right\_elbow\_y,right\_elbow\_z,right\_elbow\_v,left\_wrist\_x,left\_wrist\_y,left\_wrist\_z,left\_wrist\_v,right\_wrist\_x,right\_wrist\_y,right\_wrist\_z,right\_wrist\_v,left\_hip\_x,left\_hip\_y,left\_hip\_z,left\_hip\_v,right\_hip\_x,right\_hip\_y,right\_hip\_z,right\_hip\_v,left\_knee\_x,left\_knee\_y,left\_knee\_z,left\_knee\_v,right\_knee\_x,right\_knee\_y,right\_knee\_z,right\_knee\_v,left\_ankle\_x,left\_ankle\_y,left\_ankle\_z,left\_ankle\_v,right\_ankle\_x,right\_ankle\_y,right\_ankle\_z,right\_ankle\_v,left\_heel\_x,left\_heel\_y,left\_heel\_z,left\_heel\_v,right\_heel\_x,right\_heel\_y,right\_heel\_z,right\_heel\_v,left\_foot\_index\_x,left\_foot\_index\_y,left\_foot\_index\_z,left\_foot\_index\_v,right\_foot\_index\_x,right\_foot\_index\_y,right\_foot\_index\_z,right\_foot\_index\_v

H,0,0.79,0.59,-0.07,1.0,0.68,0.54,0.27,1.0,0.69,0.56,-0.32,1.0,0.68,0.72,0.28,0.27,0.69,0.74,-0.36,0.97,0.81,0.74,0.08,0.27,0.82,0.75,-0.18,0.92,0.46,0.54,0.19,1.0,0.47,0.55,-0.19,1.0,0.27,0.57,0.2,0.24,0.27,0.58,-0.15,0.96,0.1,0.61,0.29,0.47,0.08,0.62,-0.07,0.97,0.06,0.6,0.29,0.62,0.05,0.61,-0.08,0.96,0.1,0.72,0.19,0.69,0.08,0.72,-0.23,0.95

2.3. Basic Squat

2.3.1. Basic technique, errors description and important landmarks

A squat is a strength exercise in which the trainee lowers their hips from a standing position and then stands back up. During the descent of a squat, the hip and knee joints flex while the ankle joint dorsiflexes; conversely the hip and knee joints extend and the ankle joint plantarflexes when standing up.

There are 3 popular errors of basic squat that will be targeted in this thesis:

* Feet placement: Feet placement is extremely important in squat. The 2 feet should place in a way that the width of 2 feet is around the width of 2 shoulders.
* Knee placement: Knee placement is not only important but could be dangerous if perform incorrectly with heavy load. During the “down” stage of the exercise, the knee should be open out to wider than the feet width.

In my research and exploration, the important MediaPipe Pose landmarks for this exercise are: left shoulder, right shoulder, right hip, left hip, right knee, left knee, right ankle and left ankle.

2.3.2. Stage detection method

In contrast with other exercises, there are 2 stages when performing squat, “up” and “down” stage. Because it is important for error detection to discriminate stage of the squat, a model is trained for this task.

Videos of contributors perform proper form of a squat of this exercise are used for to build a machine learning model. Figure 13 is a visual graph represent the number of frames gathered from the videos and their classes. There are a total of 374 images, in which, there are 188 samples (50.25%) belong to class *up* and 186 (49.75%) samples belong to class *down*.

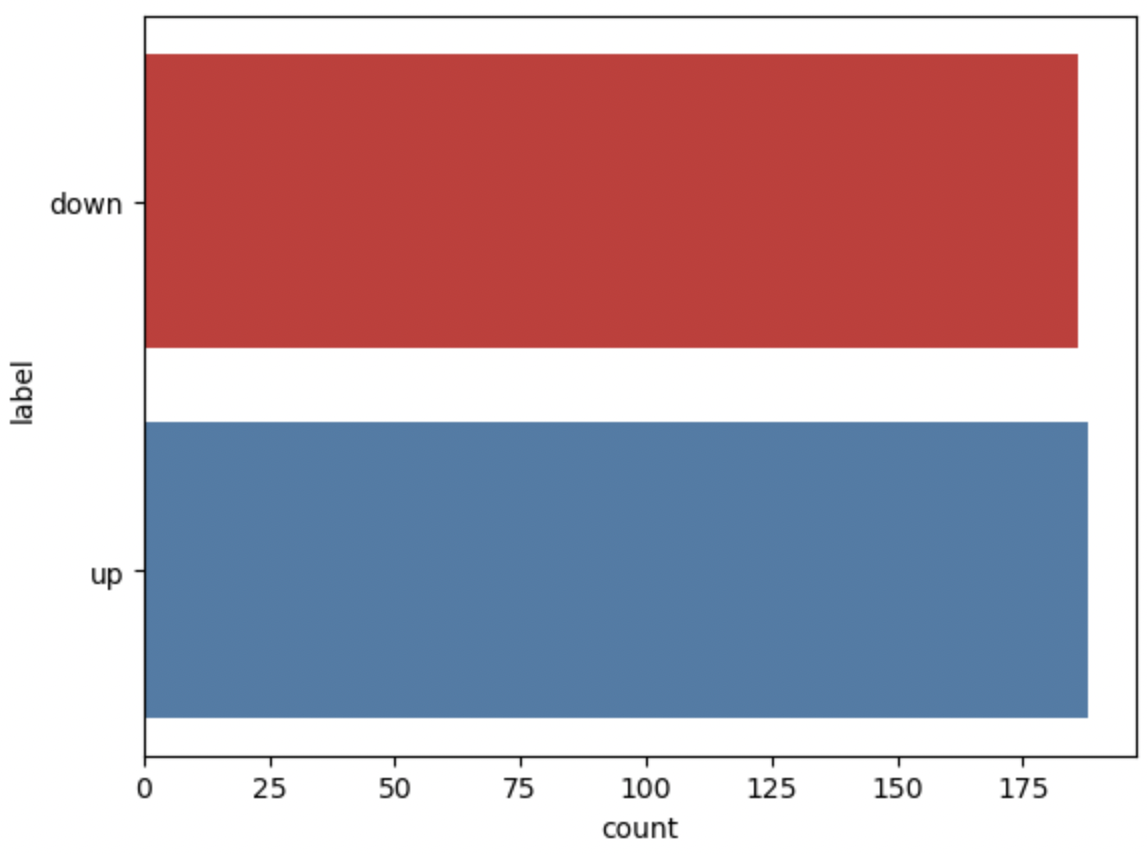


Figure 13: Class balance of Squat's dataset

Every datapoint in the dataset is consist of 37 columns which are formed based on its classed and the important landmarks. Below is an example of a datapoint:

label,,nose\_x,nose\_y,nose\_z,nose\_v,left\_shoulder\_x,left\_shoulder\_y,left\_shoulder\_z,left\_shoulder\_v,right\_shoulder\_x,right\_shoulder\_y,right\_shoulder\_z,right\_shoulder\_v,left\_hip\_x,left\_hip\_y,left\_hip\_z,left\_hip\_v,right\_hip\_x,right\_hip\_y,right\_hip\_z,right\_hip\_v,left\_knee\_x,left\_knee\_y,left\_knee\_z,left\_knee\_v,right\_knee\_x,right\_knee\_y,right\_knee\_z,right\_knee\_v,left\_ankle\_x,left\_ankle\_y,left\_ankle\_z,left\_ankle\_v,right\_ankle\_x,right\_ankle\_y,right\_ankle\_z,right\_ankle\_v

0,0.6,0.43,-0.01,1.0,0.65,0.52,0.04,0.99,0.57,0.53,0.13,1.0,0.66,0.71,-0.04,1.0,0.62,0.72,0.04,1.0,0.65,0.66,-0.3,1.0,0.54,0.65,-0.13,0.98,0.66,0.84,-0.29,0.99,0.58,0.81,-0.16,0.97

2.3.3. Errors detection method

a. Feet placement

Can be detected by calculate ratio between the distance of 2 feet and the distance of 2 shoulders. To precisely choose the correct ratio, videos of contributors perform proper form of a squat are analyzed. In that respect, 851 datapoints are gathered. Figure 14 is a visual graph represent the ratio between the feet width and shoulder width calculated from 851 data points. In conclusion, assume that is the ratio between feet width and shoulder width:

* : correct foot placement
* : foot placement is too tight
* : foot placement is too wide

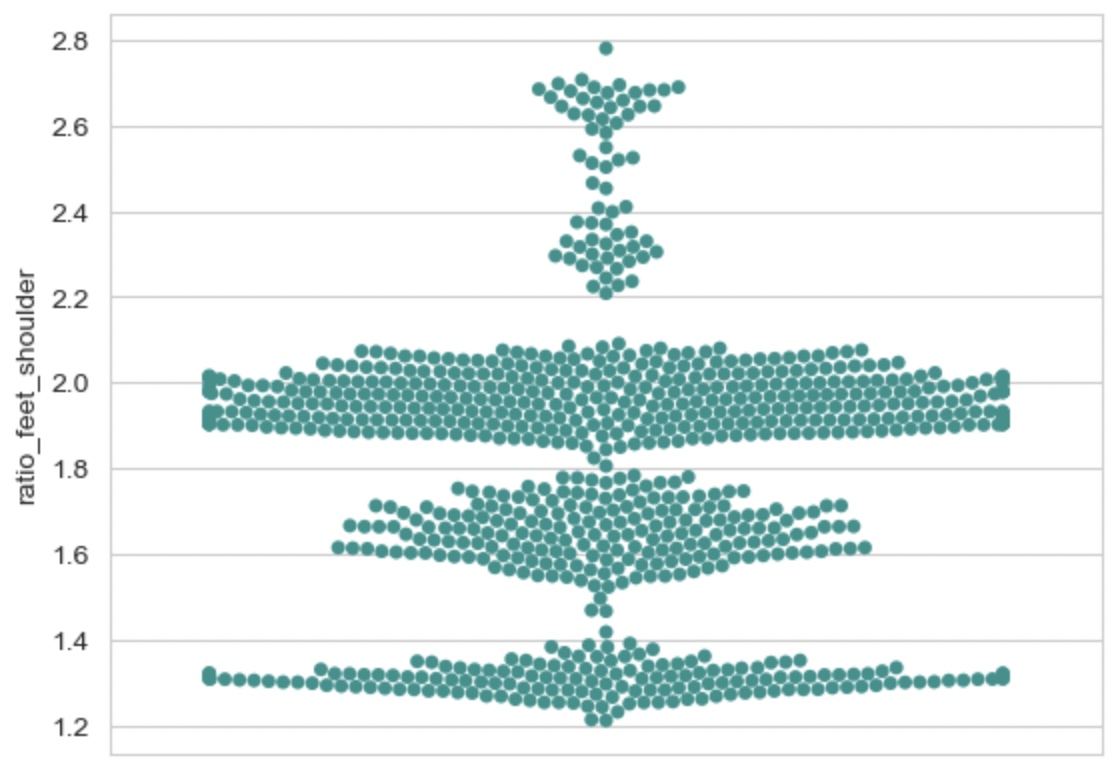


Figure 14: Graph representing the ratio between feet width and shoulder width in squat

b. Knee placement

Can be detected by calculating ratio between the distance of 2 knee and 2 feet. Similar to the previous error, videos of contributors are analyzed to determine a correct threshold. Due to the dynamic movement of the knee during the exercise, the calculated ratio from the data will be separate into 3 stages: up, middle and down.

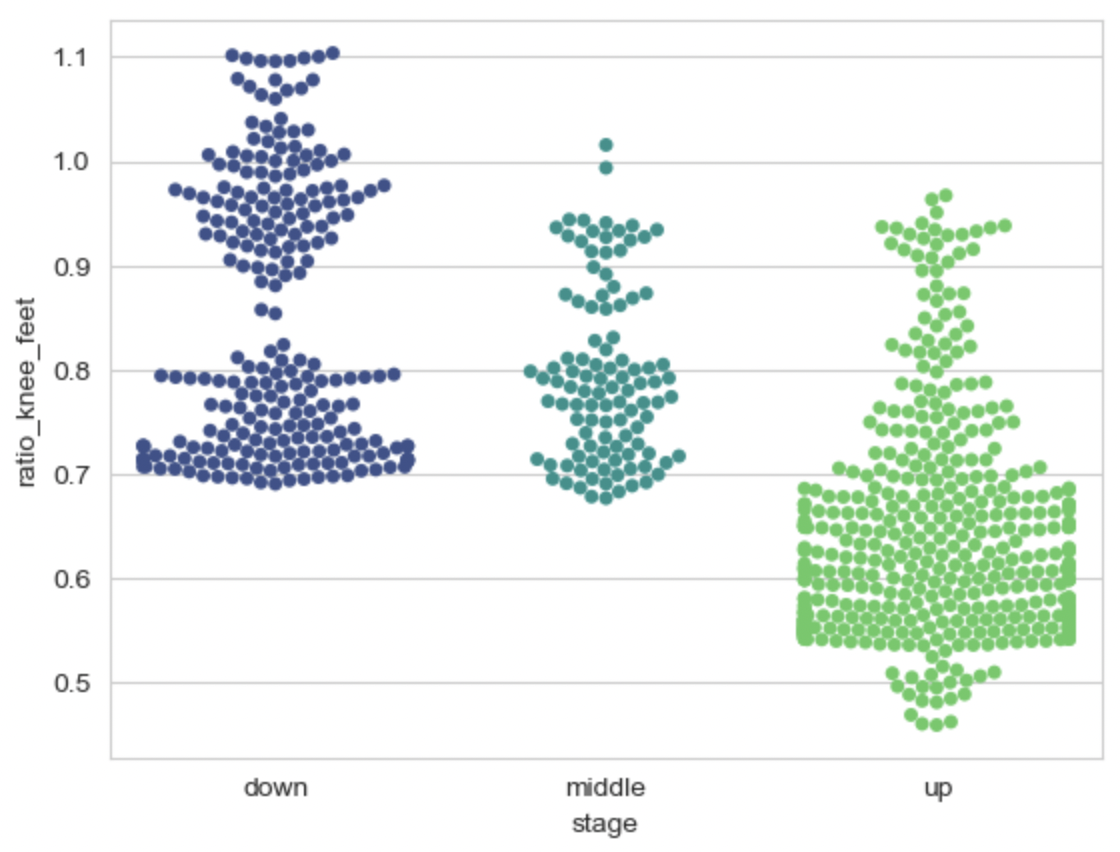


Figure 15: Graph representing the ratio between knee width and feet width in each stage of squat

2.4. Lunge

2.4.1. Basic technique, errors description and important landmarks

A lunge can refer to any position of the human body where one leg is positioned forward with knee bent and foot flat on the ground while the other leg is positioned behind.

There are 2 popular errors of lunge that are targeted in this thesis:

* Knee angle: while the body is in a low position, the angle of 2 knees should be around 90 degrees. The knee’s angle too wide or too tight could be an indication of the performer takes too big or too narrow of a step while performing lunge.
* Knee over toe: while the body is in a low position, especially with heavier weight, the performer usually leans too much forward cause the knee to go past the toe. This popular error can result in a knee injury or an imbalance for the performer.

In my research and exploration, the important MediaPipe Pose landmark for lunge are: nose, left shoulder, right shoulder, left knee, right knee, right hip, left hip, left ankle, right ankle, left heel, right heel, left foot index and right foot index.

2.4.2. Error detection methods

a. Knee angle

Can be detected by calculating the angle of the left and right knee. To precisely choose the correct lower and upper thresholds for this error, videos of contributors perform correct form of the exercise are analyzed. In that respect, 9906 datapoints of correct knee angles are gathered.

Figure 16 is a visual graph represent the angles of right and left knee while properly performing lunge. In conclusion from the graph, the angle of left/right knee during the low position of a lunge should be in between 60 and 135 degrees.

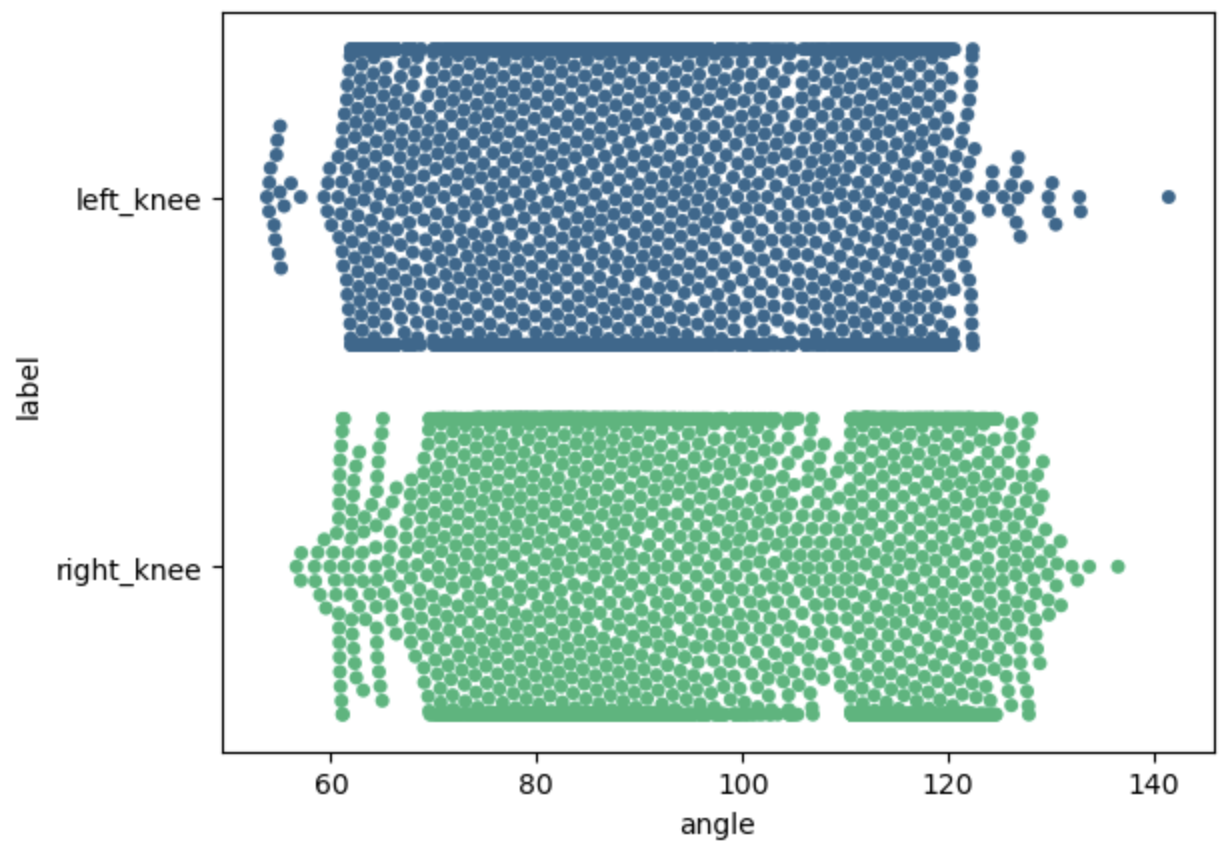


Figure 16: Graph representing the angles of right and left knee in correct lunge position

b. Knee over toe

Machine learning is implemented for this error, due to its complexity. Videos of contributors perform in both 2 stages (proper form and knee-over-toe form) of this exercise are used for to build a machine learning model. Figure 17 is a visual graph represent the number of frames gathered from the videos and their classes. There are a total of 17907 images, in which, there are 8793 samples (49.1%) belong to class *correct form (C)* and 9114 samples (51.9%) belong to class *knee-over-toe (L)*.

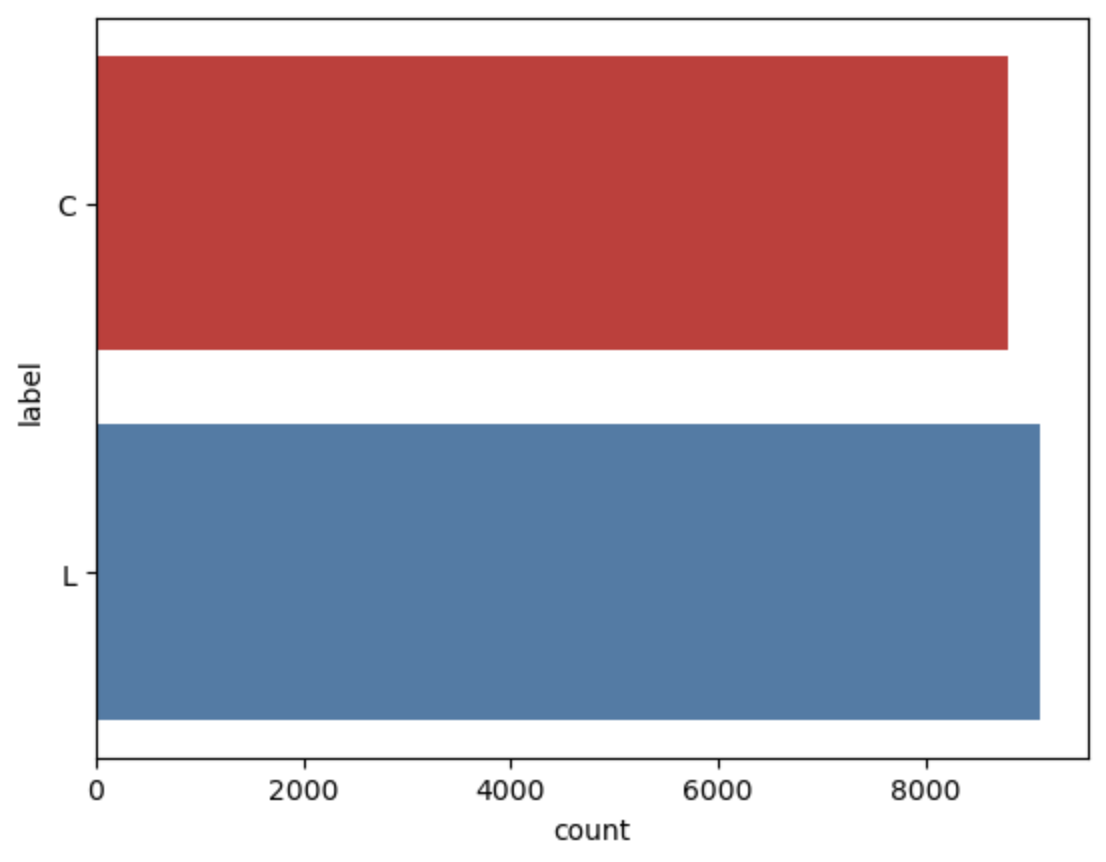


Figure 17: Class balance of Lunge’s dataset

Every datapoint in the dataset is consist of 53 columns which are formed based on its classed and the important landmarks. Below is an example of a datapoint:

label,nose\_x,nose\_y,nose\_z,nose\_v,left\_shoulder\_x,left\_shoulder\_y,left\_shoulder\_z,left\_shoulder\_v,right\_shoulder\_x,right\_shoulder\_y,right\_shoulder\_z,right\_shoulder\_v,left\_hip\_x,left\_hip\_y,left\_hip\_z,left\_hip\_v,right\_hip\_x,right\_hip\_y,right\_hip\_z,right\_hip\_v,left\_knee\_x,left\_knee\_y,left\_knee\_z,left\_knee\_v,right\_knee\_x,right\_knee\_y,right\_knee\_z,right\_knee\_v,left\_ankle\_x,left\_ankle\_y,left\_ankle\_z,left\_ankle\_v,right\_ankle\_x,right\_ankle\_y,right\_ankle\_z,right\_ankle\_v,left\_heel\_x,left\_heel\_y,left\_heel\_z,left\_heel\_v,right\_heel\_x,right\_heel\_y,right\_heel\_z,right\_heel\_v,left\_foot\_index\_x,left\_foot\_index\_y,left\_foot\_index\_z,left\_foot\_index\_v,right\_foot\_index\_x,right\_foot\_index\_y,right\_foot\_index\_z,right\_foot\_index\_v

0,0.6,0.43,0.01,1.0,0.65,0.52,0.04,0.99,0.57,0.53,0.13,1.0,0.66,0.71,-0.04,1.0,0.62,0.72,0.04,1.0,0.65,0.66,-0.3,1.0,0.54,0.65,-0.13,0.98,0.66,0.84,-0.29,0.99,0.58,0.81,-0.16,0.97 …

1. TESTING AND EXPERIMENTAL RESULTS

CONCLUSION

1. Final results

Throughout the time of research and producing this thesis with the “Delivery management system” topic, I had an opportunity to expose myself with a lot of web development technologies and services. I am also able to build a full functional delivery system that matches all the basis criteria. My system can help the manager to manage the employees, the pricing list and the orders, track their locations and the customer can place an order, track the order. The system’s user interface is built with user experiment in mind and met the goal of being user friendly, easy to use and comprehensible.

2. Future works

In this topic, due to limited time and resources I mainly focused in developing and finalizing the basis and most important features of a delivery system. In other hand, I also prepare some future paths to improve the system in the future.

* Using VISA cards to use the Google Map API for better finding and routing between locations.
* Implementing online paying methods such as PayPal, MoMo for the consignor to pay the delivery fee right after he/she confirms the order placement.
* Researching and implementing the “Return goods” policy for customers to return the products that they are not happy with.
* Developing a fully functional mobile app for company’s drivers to update their location and route their way to the consignee location.

REFERENCES

1. Giao hàng tiết kiệm, “Dịch vụ Giao hàng tiết kiệm” from <https://giaohangtietkiem.vn/dich-vu-giao-hang-ghtk>.
2. Giao hàng nhanh, “Dịch vụ giao hàng nhanh” from <https://ghn.vn/pages/dich-vu-giao-hang>.
3. Andrew Smith, “Building Management System (BMS)”, *City of Melbourne*.
4. MDN Web Docs October 8th 2021, accessed November 1st 2021, <https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django/Introduction>.
5. Rohan Vats, “Introduction To Django REST Framework: Development with Examples”, *upGrad*, 10/05/2021.
6. Vuejs.org, “Vue.js Introduction” from <https://vuejs.org/v2/guide/>.
7. Wikipedia December 5th 2021, accessed October 20th 2021, <https://en.wikipedia.org/wiki/Leaflet_(software)>.
8. Per Liedman, Leaflet Routing Machine 2015, accessed October 20th 2021, <http://www.liedman.net/leaflet-routing-machine/>.
9. Open Route Service, accessed October 20th 2021, <https://openrouteservice.org/>.