EDA Sample

Here we provide you with some samples from disparate sources, please see references for more detail or ideas.

Describe the problem

How much to write ? If your project requires more domain knowledge elaborate here for the readers vs. if your project uses IMDB / Sentiment analysis – it's ok to be brief

2.1 OVERVIEW

PRIMARY TASK DESCRIPTION

The Isolated Sign Language Recognition competition's goal is to classify isolated American Sign Language (ASL) signs. You will create a TensorFlow Lite model trained on labeled landmark data extracted using the MediaPipe Holistic Solution.

The evaluation metric for this contest is simple classification accuracy

IMPORTANT RELEVANT TERMS

- **Mediapipe:** A framework for building multimodal (eg. video, audio) applied ML pipelines. It simplifies building machine learning applications by providing a streamlined path from research prototyping to production deployment.
- **American Sign Language (ASL):** A complete, natural language that employs signs made with the hands and other movements, including facial expressions and postures of the body, used primarily by people who are deaf or hard of hearing.
- **TensorFlow Lite:** A lightweight and cross-platform framework for deploying machine learning models on mobile and embedded devices. It enables on-device machine learning inference with low latency and a small binary size.
- **PopSign:** A smartphone game app that makes learning American Sign Language fun, interactive, and accessible. Players match videos of ASL signs with bubbles containing written English words to pop them.
- **Landmark Data:** A set of labeled landmark data extracted from raw videos using the MediaPipe Holistic Solution. This dataset is used to train machine learning models for isolated American Sign Language recognition in the competition.
- **Isolated Sign Language Recognition:** The task of classifying isolated American Sign Language signs. In the competition, participants create a TensorFlow Lite model trained on the provided landmark data to recognize the signs and improve PopSign's ability to help teach ASL to parents of deaf children.

Explain the problem with visualization/images

Explain With Pictures

American Sign Language Hand Gestures in Isolation

Actual data / images /text

Brief data overview

Q: How does the data look like? What is the general feel of the numbers?

A: We have the following info:

- `site_id` - There are 2 total hospitals from where the records were gathered, split roughly 50-50
- `patient_id` - The unique identifier of the patient. There are **11.913** total patients
- `image_id` - The unique identidier of the image. There are **54.706** unique images in train. *Each patient has an average of 4.5 breast scans* (with 4 being the least number of scans and 14 being the maximum number of scans per patient).
- `laterality` - L is for the left breast, R is for the right. There are slightly more images for the R (right) breast → 27,439 than for the L (left) breast → 27,267.

Descriptive stats

```
Number of TOTAL images: 54706
Records gathered in Site 1: 29519
Records gathered in Site 2: 25187
--------------------------------------------------
Total unique patients: 11913
--------------------------------------------------
Total unique images: 54706
--------------------------------------------------
Statistics: Images per Patient
count    11913.000000
mean         4.592126
std          1.133216
min          4.000000
25%          4.000000
50%          4.000000
75%          5.000000
max         14.000000
Name: image_id, dtype: float64
--------------------------------------------------
Image records count per laterality (R): 27439
Image records count per laterality (L): 27267
--------------------------------------------------
Image records count per View:
MLO    27903
CC     26765
AT        19
LM        10
ML         8
LMO        1
Name: view, dtype: int64
```

Examine the target

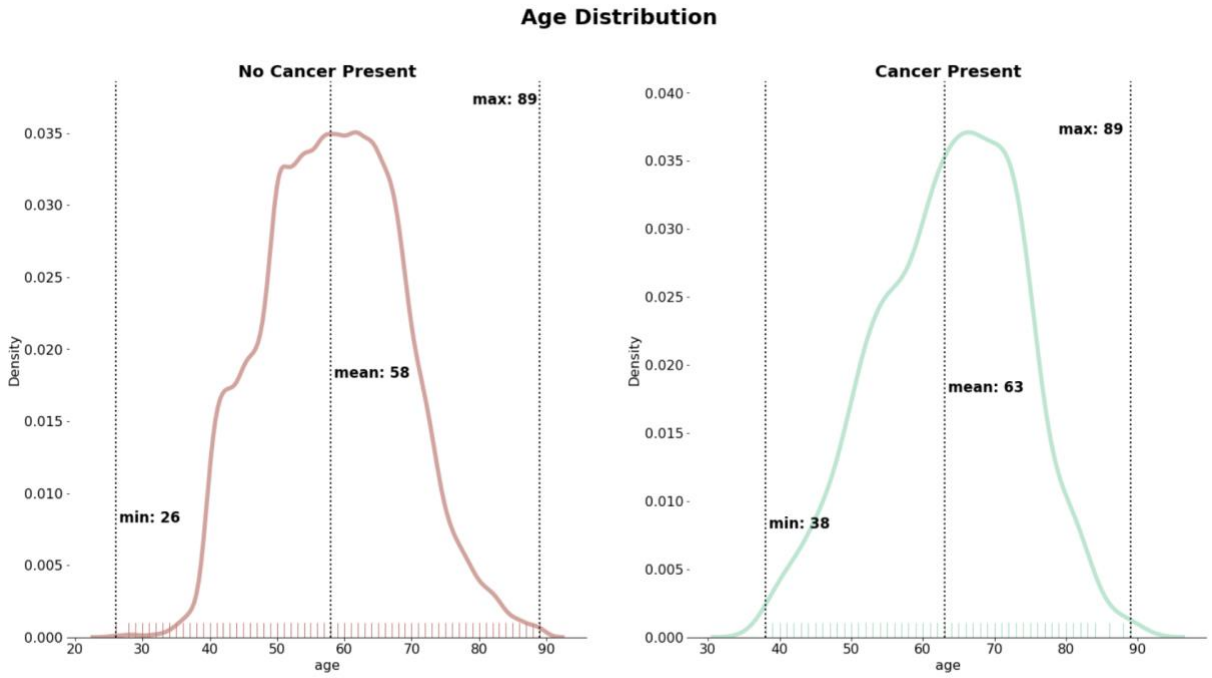
5.4 EXAMINE THE `SIGN` COLUMN

This is the label for each respective event/sequence.

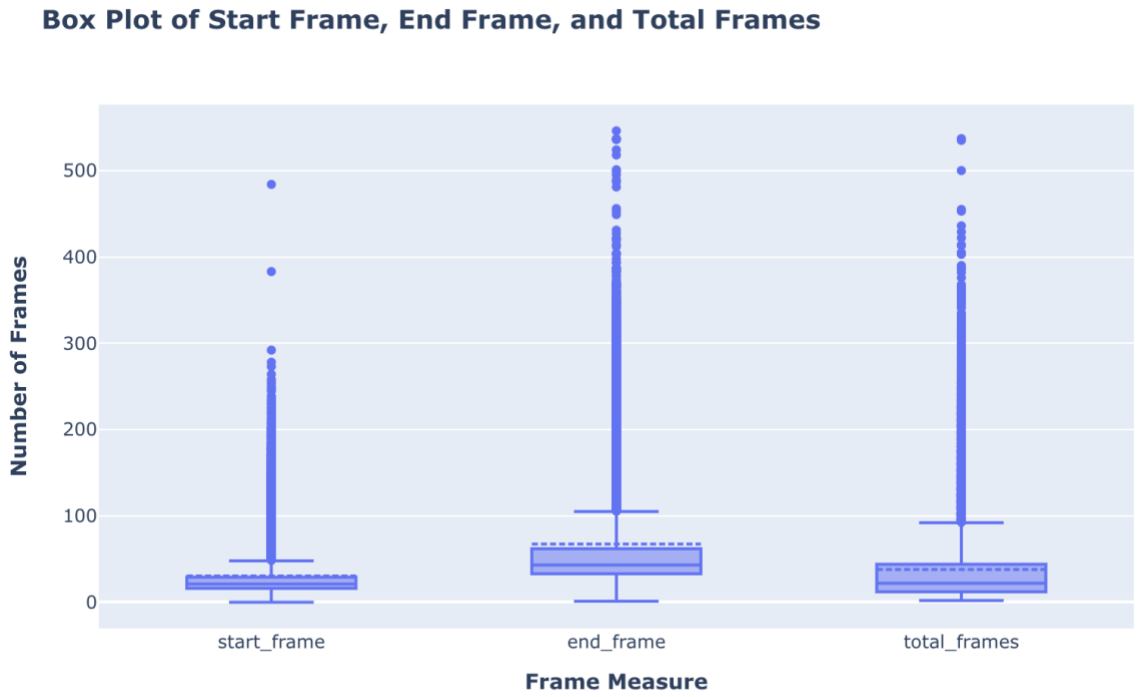
- **Number Of Unique Signs**: 250
- **Average Number of Rows Per Sign**: 377.908
- **Standard Deviation in Counts Per Sign**: 19.356537293638034
- **Minimum Number of Examples For One Sign**: 299
- **Maximum Number of Examples For One Sign**: 415

It's a pretty balanced dataset!

Distribution plots



Outlier Analysis



Insights that may be helpful during modeling phase

Quick Takeaways

- Face points can be NaN although it is less common than in the Hand data
- Pose points are never NaN
- Left and Right hand distributions are similar but Right Hand is full NaN less than Left Hand
- Pose, Left-Hand, and Right-Hand all have intermediate (not all missing or all present) sequences, however, they are less common than the case where all points are NaN or valid.

Baseline Model

Ref:
(1) https://www.kaggle.com/code/andradaolteanu/rsna-breast-cancer-eda-pytorch-baseline
(2) https://www.kaggle.com/code/dschettler8845/gislr-learn-eda-baseline