Data Challenges in Building a Facial Recognition Model and How to Mitigate Them

January | 2023

Diversity, Equity and Inclusion

Actuarial Innovation and Technology

Data Challenges in Building a Facial Recognition Model and How to Mitigate Them

AUTHOR Victoria Zhang, FSA, FCIA

SPONSORS Diversity, Equity and Inclusion Research Advisory Council and

Actuarial Innovation and Technology Strategic Research Program Steering Committee

CONTENTS

Data Challenges in Building a Facial Recognition Model and How to Mitigate Them

Introduction

Traditional underwriting process involves many manual touch points, which usually takes 4–6 weeks to complete. The recent COVID-19 lockdowns and ongoing physical-distance protocols reinforce the need to modernize the underwriting process. Life insurance companies need to build a digitally enabled, data-augmented, life product purchasing journey. Cutting-edge artificial intelligence (AI) technology has been studied to replace traditional underwriting procedures. In recent years, the life insurance industry has started exploring AI technologies such as facial recognition for various purposes.

In 2017, a United States-based startup, Lapetus Solutions, released a disruptive product named JANUS which can estimate a person's age, biological sex, and BMI from the analysis of a simple selfie. Lapetus has other products such as CHRONOS which will give an underwriting score of a person by a selfie and a few simple questions, which they claim can be more accurate than the traditional blood testing during underwriting. Such underwriting process usually takes less than 10 minutes and makes the life insurance product purchase a lot easier to purchase.[1]

Outside of North America, Ping An—the biggest insurance company in China—uses facial recognition technology to make sure the customers are not sick during underwriting. "The company claims that it can assess risks with greater accuracy and reduce fraud. Its technology is reportedly the best available, with an accuracy rate of 99.8 percent."[2] Facial recognition technology can detect emotions and, in particular, suspicious behavior. It can do this by monitoring micro expressions, eye movement, pupil dilation, gaze, speech variations and tone. It is also possible to identify seven main human emotions.

While many people are enjoying the positive outcomes of facial recognition, there are some caveats that cannot go unnoticed. Facial recognition is one of the computer vision technologies based on the recent success of deep learning models such as convolutional neural networks. However, some lawmakers have raised serious concerns about the use of facial recognition, especially in the context of racial disparities.

A successful machine learning model requires large amounts of data. "Garbage in, garbage out" is a cliché in the machine learning area, which means that the quality of the input training data plays a vital role in the quality of the machine learning model results. Training facial recognition requires a lot of facial image data, and if the training image data has imbalanced samples, the prediction accuracy of the trained model will often lead to bias.

Before embracing AI technology to simplify the insurance underwriting process, the author wants to bring actuaries on board to understand how the technology works and what potential risks it could introduce. The author wants to focus on data bias as it is one of the main caveats of facial recognition technology. This paper would like to assess some publicly available dataset to test how data homogeneity could affect facial recognition models. The intention

[1] Lapetus Solutions Inc, "How JANUS Works," https://www.lapetussolutions.com/solutions/janus/, Accessed 14 December 2022
[2] Enoch You, 6 April 2018, "How Chinese Insurer Ping An Uses Facial Recognition to stop Its Agents Skipping Meetings," South China Morning Post, https://www.scmp.com/business/companies/article/2140409/how-chinese-insurer-ping-uses-facial-recognition-stop-its-agents

of this project is to close the gap about such bias understanding, and then to try to explain and analyze such bias from the actuarial point of view.

The report will be structured as the following:

1. Describe how facial recognition models work in general
2. Discuss and analysis several publicly available datasets for facial recognition models
3. Implement a facial recognition model
4. Train and test the model based on the public datasets
5. Discuss potential bias introduced and propose methodologies that could fix or limit the bias

What is Facial Recognition?

Facial recognition is a technique to identify the face of individuals from photos, and this technique has been widely used in many areas such as security, law enforcement, access control, immigration, health care, etc. This technique now is used everywhere in our daily lives. For example, big tech companies such as Apple and Google leverage facial recognition in their mobile devices for user identity verification.

Historically , facial recognition is a technology that works by identifying a person's face to perform actions such as unlocking a phone. It works by calculating the nodal points of a person's face. Basically, it calculates the distances between prominent features (eyes, nose, etc.) and uses those points as coordinates to create a face map. A person typically has 80 nodal points. Once the coordinates are gathered, the operating system creates a faceprint based on the information. The software pays special attention to specific measurements such as the distance between eyes and the width of cheekbones. The information is stored in a database so that it can be used later to verify the user.

The actual technology behind the facial recognition varies from one application to another; however the basic mechanics are similar and usually involve the following basic steps:

- **Step 1:** Detect the face region from the image.
- **Step 2:** Extract the geometry of the face. Every face has some various and visible landmarks called nodal points. The different peaks and valleys are forming geometry features. A human face has roughly 80 nodal points, which are measured by computer software. There is also other geometry information such as distance between eyes, width of the nose, distance between nose and mouth, distance between eyes and tip of the nose, depth of the eye sockets, etc.
- **Step 3:** Create faceprint from the geometry information (Figure 1), which will be used to query the existing database to identify the individual.

Figure 1

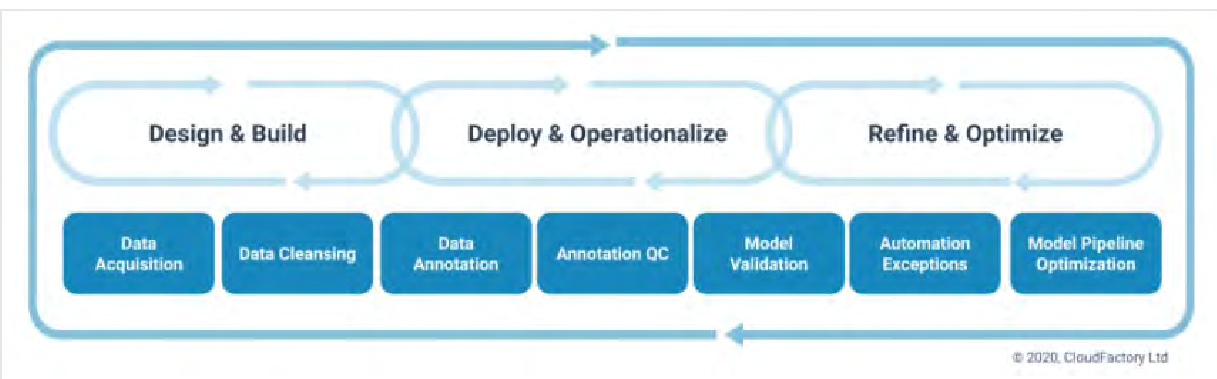
Broadly speaking, facial recognition can also refer to techniques that extract information such as healthiness, age, sex, race and so on for various purposes. Thanks to progress in machine learning and deep learning, facial recognition technology is rapidly maturing. Especially because of advancement of deep learning techniques in the computer vision area, learning-based facial recognition techniques have achieved remarkably accurate results, and in certain usages, it outperforms humans. Therefore, the focus of this report will be the learning-based method.

Learning-Based Facial Recognition Technique

Figure 2 is a graph from CloudFactory explaining the process of a learning-based model development, which also applies to learning-based facial recognition models. To develop a new learning-based model, a high-quality training dataset is the key. The first thing we need is to gather a fairly large amount of training data.

Figure 2

LEARNING-BASED MODEL DEVELOPMENT



For example, if one wants to design a facial recognition technique to predict body mass index (BMI) from only a facial image, one will need to gather facial images with ground truth BMI information for those images. In the machine learning area, the facial images can be denoted as X, and Y is usually used to denote ground truth BMI information. The goal is to design a model with up to millions of to-be-determined parameters, which the author denotes as $\hat{Y} = f_{\theta}(X)$, and $\hat{Y}$ is a prediction of the provided ground truth label Y by adjusting the model parameters Θ with the gradient descent methods.

Mathematically, the training process can be summarized as this equation $\hat{\Theta} = argmin \, || Y - f_{\Theta}(X) ||$. The trained model is $f_{\hat{\Theta}}(.)$, which will be used in the deployment to predict the BMI information of this image. It is worth mentioning that that the dimension of the model parameters Θ can be up to a few million, and thus many images (at least a few thousand) and their labels are required to train the model.

Training data is the cornerstone of facial recognition models. In the following sections the author will look closely at a few public facial recognition datasets and explore the statistics distribution of those datasets.

Facial Recognition Public Datasets

This section will explore a few popular and public facial recognition datasets that are widely used as the training facial recognition machine learning models.

IMDB-WIKI

IMDB-WIKI is one of the largest publicly available datasets of facial images with both sex and age labels.[3] Celebrities' profiles were picked from the IMDB or Wikipedia websites with their biological ages and pictures. Assuming that the images with single faces have the correct timestamp and link the correct date of birth, I can assign the true biological age of each image.

There are a total 209,990 images downloaded, and I did a high-level assessment of the dataset:

1. Sex distribution (Figure 3) is relatively even, about 42% female and 58% male.
2. Skewness in age distribution (Figure 4) appeared between age 20 to age 40, comprising 58% of the total population.

Unfortunately, IMDB-Wiki dataset does not have a label for race or ethnicity, so the author is unable to analyze race or ethnicity information.

[3] Rasmus Rothe, Radu Timofte, and Luc Van Gool, 2018, "Deep Expectation of Real and Apparent Age From a Single Image Without Facial Landmarks," International Journal of Computer Vision, Volume 126, No. 2–4, pp 144–157, https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

Figure 3

IMDB-WIKI SEX DISTRIBUTION

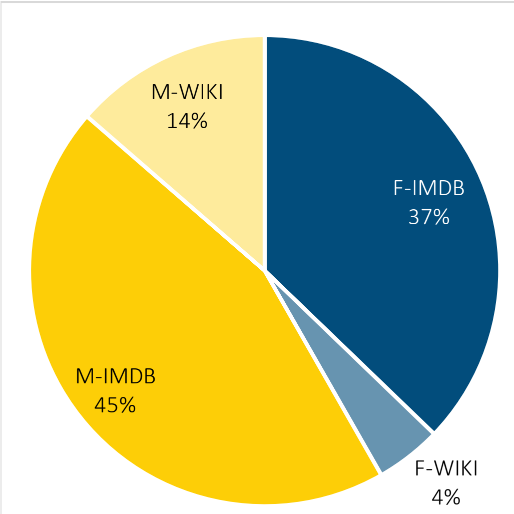
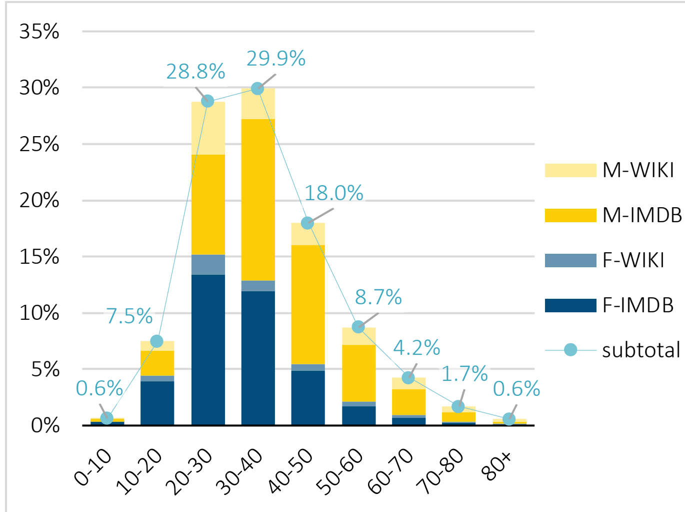


Figure 4

IMDB-WIKI POPULATION DISTRIBUTION BY AGE GROUP



UTK FACE

UTK Face dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, sex, and ethnicity. The images cover large variation in pose, facial expression, illumination, resolution, etc. The dataset is a very popular choice for training and testing facial recognition model.

Figure 5 is a sample images demonstrates how the images look like in this UTK dataset.

Figure 5
SAMPLE UTK DATASET IMAGES

First the author analyzes the sample age distribution (Figure 6), and it is quite noticeable that there are more samples with age less than or equal to 10 years old, which accounts for 32% of the whole dataset. This indicates that this dataset is slightly imbalanced and has disproportionately more samples of children and infants. The remaining age groups are relatively evenly distributed, except for a decreasing ratio trend as age increases.

Figure 6
UTK FACE DATA DISTRIBUTION BY AGE GROUP

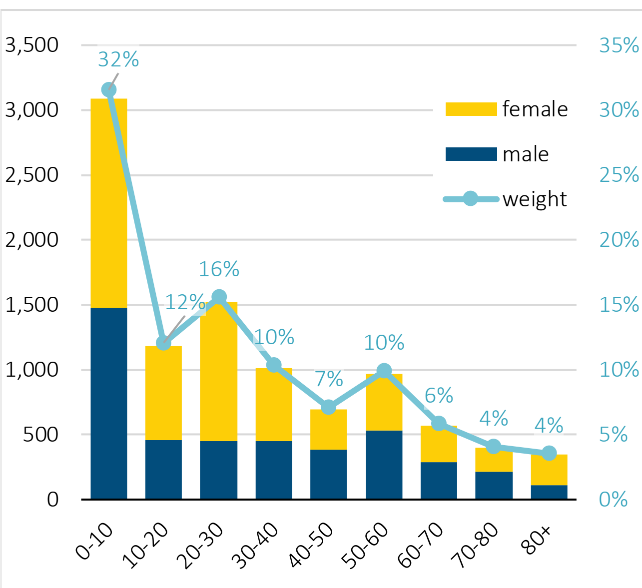
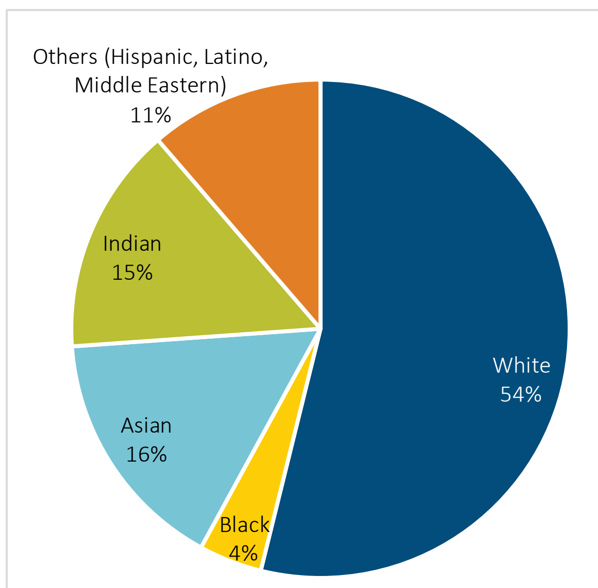


Figure 7
UTK RACE DISTRIBUTION VISUALIZATION



As for the sex distribution, 55.3% of samples are female, and 44.7% of samples are male. The imbalance is not severe.

The author finds the racial/ethnic group distribution (Figure 7) to be the most interesting part of the UTK Face dataset analysis. One can clearly observe that the number of images in the white racial/ethnic group is greater than the combined total samples of all other racial/ethnic groups.

APPA-REAL DATABASE

The APPA-REAL database contains 7,591 images with associated real and apparent age labels. The apparent age labels are crated by human eyes, with on average around 38 votes collected for each image to get the average perceived apparent age. APPA-REAL is an ideal dataset for us to compare the performance of a machine learning model with human eye judgement.

In the following sections the author will introduce how to address the imbalanced datasets from a model training point of view to avoid biased results. The author will discuss how to build a deep learning model for facial recognition for age and sex prediction, leveraging the three public datasets discussed above to train and evaluate the deep learning model.

Case Study—Age and Sex Estimation

The author will perform a case study in this section and build a deep learning model that is trained with the IMDB-Wiki dataset to predict age and sex based on facial images.

The trained model will predict biological sex and age of each testing profile picture, and the author will analyze the testing results. The author will test its performance against the apparent age labels of the APPA-REAL database. To demonstrate that using an imbalanced dataset might cause prediction bias, the author will evaluate whether potential bias exists within the training dataset, which has an imbalanced data distribution between different age group, via UTK Face data.

Questions the author will try to address from UTK Face testing results:

- Given the uneven distribution between different **age group** of the training data, will performance differ from testing results?
- What about the model performance by **sex** group?

AGE AND SEX ESTIMATION WITH DEEP LEARNING

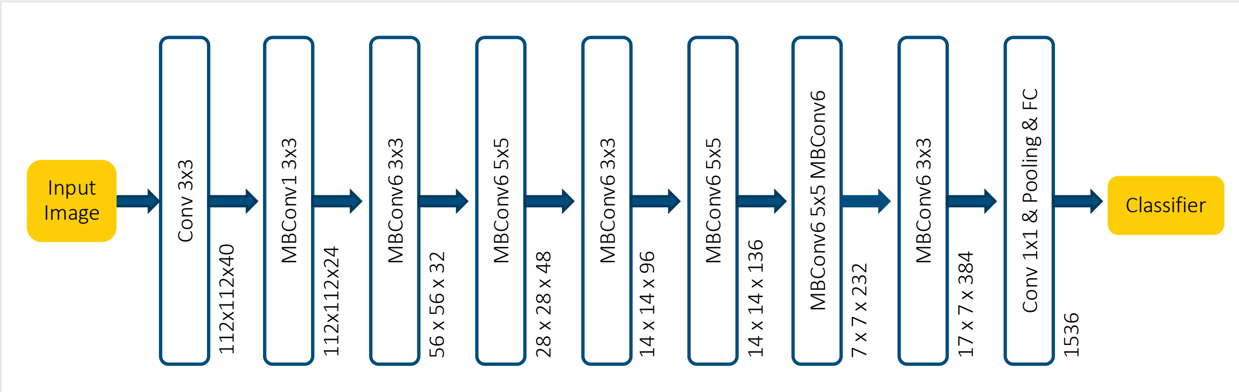
EFFICIENTNET MODEL

The model I used is called EfficientNet, which is a lightweight Convolutional Neural Network (CNN). We are using EfficientNet-B3, which has about 12 million parameters [https: //arxiv.org/pdf/1905.11946.pdf.](https://arxiv.org/pdf/1905.11946.pdf)

The schematic of the EfficientNet-B3 looks like Figure 8 where Conv denotes one block of convolutional neural layers, and the author models the age and sex estimation as a classification problem.

Figure 8

EFFICIENTNET-B3 SCHEMATIC



AGE-SEX ESTIMATION

The model used here for age and sex estimation is built on top of EfficientNet-B3.Its high-level architecture is illustrated in Figure 9 and has three processing modules:

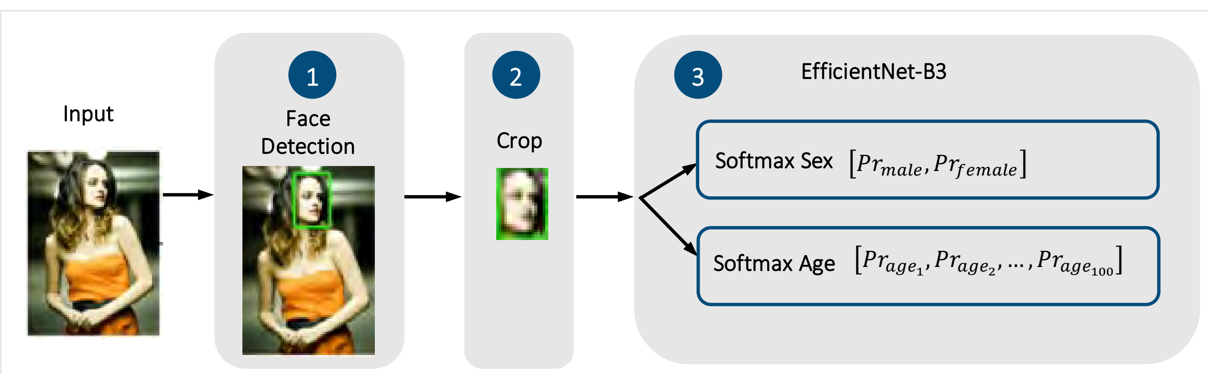
1. Face Detection: the face detection module will detect a bounding box from an input image. In this report, a dlib face detection algorithm is used to detect the face.

2. Crop: crop the original input image to the bounding box. The cropped image will then be fed into the age and sex estimation module.

3. Estimate: EfficientNet-B3 module will estimate age and sex from the cropped facial image. It is worthwhile to mention that one can replace the EfficientNet-B3 with other heavy weight network to achieve even better accuracy. The author chose this network purely to limit the training time to be under 48 hours with a consumer grade GPU such as NVIDIA GeForce 1080.

Figure 9

AGE-SEX ESTIMATION



Sex prediction is achieved by using the softmax function to predict the probability of male Pr_{male} and female Pr_{female}. Note that $Pr_{male} + Pr_{female}$ = 1.0. One can map the probability to sex simply by:

$$Predicted\ Sex = \begin{cases} Male & if\ Pr_{male} \geq 0.5 \\ Female & otherwise \end{cases}$$

Age prediction is also performed with the help of softmax. The author models the age prediction as a classification problem instead of regression problem, where the author is trying to decide which age from age 1 to 100 best fits the input facial image. Again, the softmax function will output one probability Pr_{age_i} for each age i, and due to the nature of the softmax funciton, where $\sum_{i=1}^{100} Pr_{age_i} = 1.$ Finally, the predicted age can be calculated as:

$$Predicted\ Age = \sum_{i=1}^{N} i \times Pr_{age_i}$$

IMDB-WIKI FACIAL RECOGNITION MODEL

The author trained the above model with IMDB datasets which have 171,853 training images with age and sex labels. The author set the training epochs to 30, which means repeating the whole training dataset 30 times. The batch size is set to 32, which means feeding the model with 32 samples at one shot together with their age and sex labels. Note that those hypo parameters are selected empirically. A very high number of epochs could improve accuracy but will increase the risk of overfitting the training dataset. A very low number of epochs will increase the risk of underfitting the training dataset. Table 1 shows the full configuration used in the researcher's training for others to reproduce the results. With NVIDIA GeForce 1080 GPU, the training takes about 48 hours.

Figure 10 shows model accuracy for age and sex prediction from each epoch. Sex prediction accuracy quickly converged to be as high as 90% after 20 epochs.

Age accuracy, defined here as age prediction equals actual age, is not as high as desired (Figure 11). It reaches only about 16% accuracy after 30 epochs because compared to sex, age is harder to study and predict in nature.

Table 1
AUTHOR'S IMDB DATASET CONFIGURATION

Parameter	Specification
Dataset	IMDB
Model	EfficientNetB3
Image Size	224x224
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	30

Figure 10
SEX PREDICTION ACCURACY

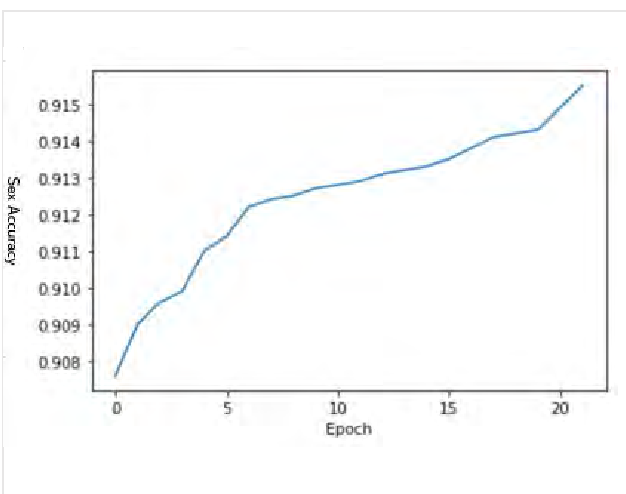
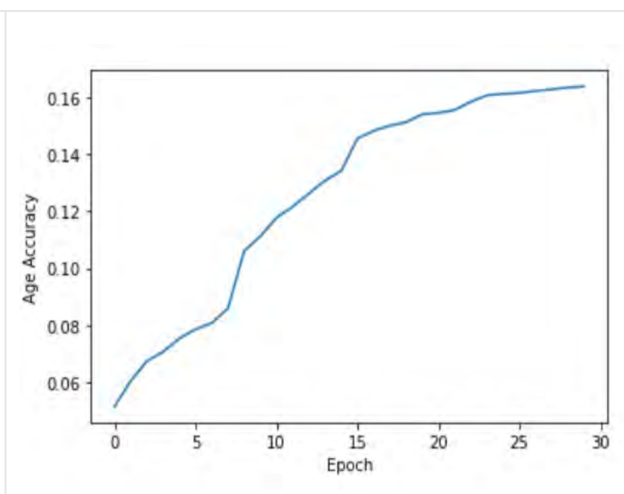


Figure 11
AGE PREDICTION ACCURACY



MODEL TESTING

To assess the impact of 16% accuracy of age estimation, the author decided to evaluate the model accuracy with a different dataset that was not used to train the author's model. The APPA-REAL database contains 7,591 images with associated real and apparent age labels. The total number of apparent votes from human eyes is around 250,000. On average, each image has around 38 votes, therefore the apparent age should be very stable. The author is using Mean Average Error (MAE) as the metric to evaluate model accuracy. Mathematically, MAE is defined as the following, and the author wants MAE to be as close to zero as possible:

$$MAE\_Apparent = \sum_{i=1}^{N} \frac{|pred\_age_i - Apparent\_age_i|}{N}$$

$$MAE\_Real = \sum_{i=1}^{N} \frac{|pred\_age_i - Real\_age_i|}{N}$$

It can be observed that the MAE for both apparent and real ages are around 5 to 6 years, which is reasonably accurate:

MAE Apparent:	5.38 years
MAE Real:	6.31 years

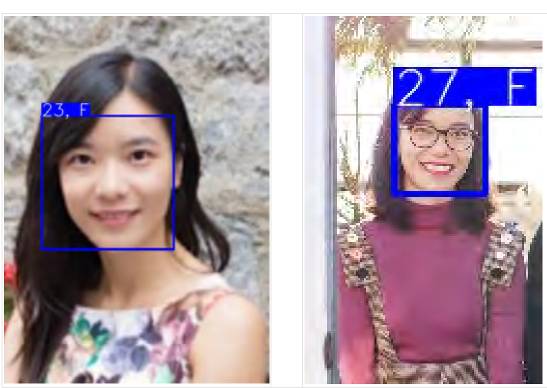
It is interesting that the MAE for apparent age is about 1 year less than the MAE for the real age. Recall that the apparent age is what a human will think the age is based on prior experience. Like human eyes, the model will guess apparent age more accurately than it will guess real age.

TEST WITH A REAL PERSON

To verify the model accuracy, the author also applies the trained model to pictures of the author at different ages. On the left in Figure 12 is a picture of the author at 26 years, and on the right she is 32 years. The trained model can correctly recognize that the sex is female, and the age prediction is also reasonably accurate (age 23 on the left and 27 on the right). What is important is that the model can predict that the person on the right is older than the person on the left.

Figure 12

AUTHOR, AGE 26 AUTHOR, AGE 32



MODEL BIAS EVALUATION

In this section, the author will use the UTK-Face dataset to evaluate the variance of the model that was trained with an imbalanced dataset. The author will evaluate the model accuracy (MAE) for age and sex. Table 2 summarizes the testing results by different age groups. The MAE_Sex column is the average sex prediction accuracy, while MAE_Age is the average age difference between predicted and real ages within the age cohort.

Model Accuracy by Age

Recall that the author's training dataset (IMDB-Wiki) is heavily skewed (over 75% data) toward the age 20–50 group. With more training data in the age groups 21–40 and 41–60, the author noticed the model performs best at ages 20–60. MAE_Sex of ages 41–60 reaches global minimum of 8.9%, meaning the model predicts correct biological sex more than 90 times out of 100 for ages 41–60.

Age is usually very tricky for a deep learning model to predict. With the author's IMDB-Wiki model, the author sees an MAE of 21.70 for people younger than 20 years[4] (the average number of years by which the model was wrong is greater than the maximum age of the cohort) and 30.44 for people older than 80 years. The training data has very limited data points (less than 10%) in these two age ranges, and the author thinks this could directly impact the prediction accuracy for these two age groups. On the flip side, for age group 21–40, the MAE of predicted age is as low as 6.69. The testing results clearly show that imbalanced training data distribution could result in uneven performance (bias) by the same facial recognition model.

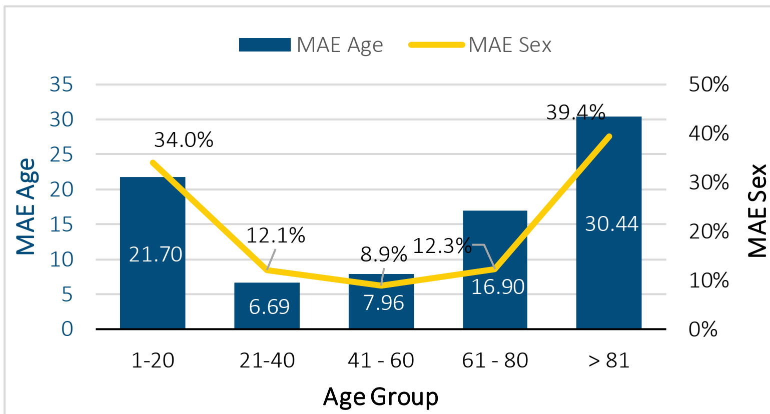
Table 2

AGE PREDICTION TESTING RESULTS

Age Group	MAE Sex	MAE Age
1–20	34.0%	21.70
21–40	12.1%	6.69
41–60	8.9%	7.96
61–80	12.3%	16.90
≥ 81	39.4%	30.44

Figure 13

BIAS EVALUATION—MAE BY AGE GROUP



Model Accuracy by Sex

The training dataset has a relatively even split between males and females, and testing results (Table 3) show the same age prediction performance for both sexes (both have MAE of 11.70).

Table 3

SEX PREDICTION TESTING RESULTS

Sex	MAE Sex	MAE Age
Male	22.0%	11.70
Female	12.0%	11.10

[4] The model's track record for estimating ages of people under age 20 is so poor that the average number of years by which the model was wrong was greater than the oldest age in the cohort.

Techniques to Mitigate Dataset Bias

The previous sections have shown the impact of imbalanced datasets, and the author has observed that it is quite common to have imbalanced datasets in the facial recognition domain.

This section will introduce a few widely used techniques for addressing imbalanced datasets. To facilitate the following discussion, assume there is a facial recognition dataset with a total of 1,000 images of different individuals, and each image is labeled either sick or healthy. One can imagine this dataset could be used in underwriting, where the software is trying to predict healthiness based on facial images. This imagined dataset is highly imbalanced and has 900 images labeled healthy and 100 images labeled sick (Table 4). Because the healthy sample is significantly larger than the sick sample, the author named it as the major class and accordingly named the sick samples as the minor class.

Table 4
SUMMARY OF IMAGINARY SAMPLES

Label	Number of Samples
Healthy	900
Sick	100

RE-SAMPLING TECHNIQUE

Re-sampling is an intuitive, simple and quite effective method to combat imbalanced datasets. To create a more balanced dataset out of the highly imbalanced one, the author either under samples or oversamples.

UNDER SAMPLING

With under sampling, one reduces the number of samples in the major class by random selection. For example, for the healthy/sick dataset, the author randomly selects 100 samples from the original 900 healthy samples, and together with the original sick samples, creates a balanced dataset with a total of 200 samples (Table 5). The issue of this sampling method is that only 200 samples remain out of the original 1,000 samples.

Table 5
SUMMARY OF IMAGINARY DATASET AFTER UNDER SAMPLING

Label	Number of Samples
Healthy	100 (After Under Sampling)
Sick	100

OVER SAMPLING

With over sampling, the author will repeat the minor class samples by multiple times to make the total of minor class samples roughly the same as the major class number of samples. With this method, the imbalanced imaginary dataset will total 1,800 samples (Table 6) because the author has increased the number of sick samples from 100 to 900.

Table 6
SUMMARY OF IMAGINARY DATASET AFTER OVER SAMPLING

Label	Number of Samples
Healthy	900
Sick	900 (After Over Sampling)

It is worth calling out that there are drawbacks to oversampling. Oversampling creates a larger but also unnaturally homogenous sick population because the same 100 individuals are repeated multiple times in the dataset.

CLASS WEIGHT TRICK

Another popular way of dealing with imbalanced datasets is applying class weights. This is an effective method to deal with imbalanced datasets. Using the fake dataset as an example, the author will first calculate a weight for each class based on the following formula, where N_i is the number of samples for class i:

$$class\_weight_i = \frac{1/N_i}{N/2.0}, \quad N = \sum_i N_i$$

The class weights will be used to weight the cost for each class, therefore the class with fewer samples will have greater weight (Table 7). This method is very similar to the oversampling method; however, it will be more computationally efficient because the author does not replicate the minority class samples.

Table 7
SUMMARY OF IMAGINARY DATASET AFTER CLASS WEIGHTING

Label	Number of Samples	Class Weight
Healthy	900	2.22×10^{-6}
Sick	100	2×10^{-5}

Conclusions

Deep learning models in facial recognition have been attracting more attention in the past years from various fields. For insurance underwriting, this technology is under exploration. The author believes facial recognition techniques will help the insurance industry by speeding up the underwriting process and fraud detections, as well as other business experience improvements. However, there are challenges in building an effective, accurate, and unbiased facial recognition model, and the author is aware that some lawmakers have raised serious concerns about the use of facial recognition.

The first part of this research walked through the fundamentals behind a facial recognition algorithm, assessed some of the most popular publicly available datasets, and conducted a case study to build a deep learning model with a dataset. The second part of the research highlighted how an imbalanced dataset could lead to uneven testing results. The report concluded with techniques to potentially mitigate bias in a dataset.

Acknowledgments

The researcher's deepest gratitude goes to those without whose efforts this project could not have come to fruition: the Project Oversight Group and others who contributed valuable ideas, thoughtful guidance and diligent review throughout the development of this research.

Project Oversight Group members:

Ann Weber, JD

Nick Gabriele, FSA, MAAA

Noel Harewood, FSA, MAAA

Mengting Kim, FSA, CERA, MAAA

Hezhong (Mark) Ma, FSA, MAAA

Mark Sayre, FSA, CERA

Shisheng (Rose) Qian, ASA, CERA

At the Society of Actuaries Research Institute:

Lisa Schilling, FSA, EA, FCA, MAAA

References

Ping An Group. 12 May 2017. "Ping An Technology Facial Recognition Technology Tops the World in LFW Test Results." https://group.pingan.com/media/news/News-2017/Ping-An-Technology-Facial-Recognition-Technology-Tops-the-World-in-LFW-Test-Results.html.

Eggar Forrester Creative. 30 June 2020. "How Facial Recognition Technology Can Be Used In Underwriting." https://www.eggarforresterinsurance.com/creative/blog/facial-recognition-technology-underwriting/.

Y. Oh, J. See, A. C. L. Ngo, R. C. Phan, and V. M. Baskaran. 15 June 2018. ''A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods and Challenges.'' arXiv:1806.05781. https://arxiv.org/abs/1806.05781.

About The Society of Actuaries Research Institute

Serving as the research arm of the Society of Actuaries (SOA), the SOA Research Institute provides objective, data-driven research bringing together tried and true practices and future-focused approaches to address societal challenges and your business needs. The Institute provides trusted knowledge, extensive experience and new technologies to help effectively identify, predict and manage risks.

Representing the thousands of actuaries who help conduct critical research, the SOA Research Institute provides clarity and solutions on risks and societal challenges. The Institute connects actuaries, academics, employers, the insurance industry, regulators, research partners, foundations and research institutions, sponsors and non-governmental organizations, building an effective network which provides support, knowledge and expertise regarding the management of risk to benefit the industry and the public.

Managed by experienced actuaries and research experts from a broad range of industries, the SOA Research Institute creates, funds, develops and distributes research to elevate actuaries as leaders in measuring and managing risk. These efforts include studies, essay collections, webcasts, research papers, survey reports, and original research on topics impacting society.

Harnessing its peer-reviewed research, leading-edge technologies, new data tools and innovative practices, the Institute seeks to understand the underlying causes of risk and the possible outcomes. The Institute develops objective research spanning a variety of topics with its strategic research programs: aging and retirement; actuarial innovation and technology; mortality and longevity; diversity, equity and inclusion; health care cost trends; and catastrophe and climate risk. The Institute has a large volume of topical research available, including an expanding collection of international and market-specific research, experience studies, models and timely research.

Society of Actuaries Research Institute
475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
www.SOA.org