



An ML-driven solution for detecting real vs. AI-generated faces in images

# Our team



**Ghiwa Lamah**

Product Manager

*EDA, Explainability*



**Nancy Ding**

User Experience

*Market & User Research*



**Melissa Olivera**

Data Expert

*User Testing*



**Jacob Petrisko**

ML Expert

*Model Evaluation*



**Queen Tran**

Architecture Lead

*EDA*

Which of these images is of a **REAL** person?



Photo 1

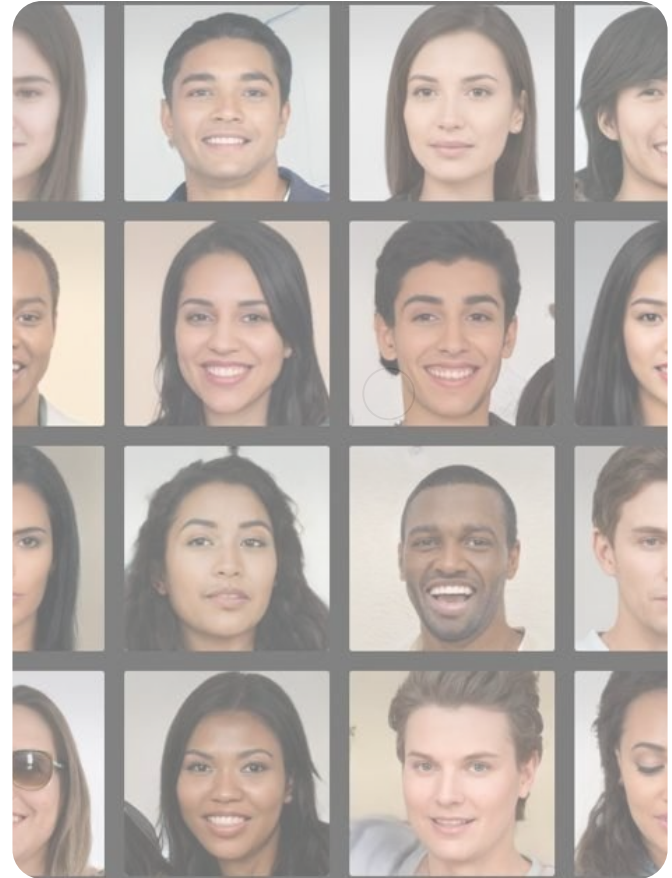


Photo 2

# Problem

With advancements in generative AI technology, it is increasingly easy to generate extremely realistic images of people who don't exist or **fake images of people** who do exist.

**90%** online content that is estimated to be AI-generated by 2026<sub>1</sub>





 **AuthentiFACE**

Facial authenticity classification tool that can label AI-generated images of people to provide more transparency to consumers of internet content

# AuthentiFace Target Users



## Social Media

Social Media Platforms

*Facebook, Instagram, Twitter, ...*



## Digital Advertising

Digital Advertising Platforms

*Google, Facebook Ads, ...*



## Dating Apps

Dating App Platforms

*Hinge, Bumble, ...*

# High Market Impact



## Social Media

**5 Billion**

*social media users worldwide in Jan 2024<sup>2</sup>*

**\$1.4 Billion**

*reported losses due to fraud on social media in the US in 2023<sup>5</sup>*



## Digital Advertising

**\$740.3 Billion**

*projected ad spending on digital advertising by 2024<sup>3</sup>*

**\$84 Billion**

*estimated cost of ad fraud worldwide in 2023<sup>6</sup>*



## Dating Apps

**\$3.12 Billion**

*projected revenue of online dating market by 2024<sup>4</sup>*

**\$1.1 Billion**

*reported losses due to romance scams in the US in 2023<sup>7</sup>*

2. Source: <https://www.statista.com/statistics/617136/digital-population-worldwide/>

3. Source: <https://www.statista.com/outlook/dmo/digital-advertising/worldwide>

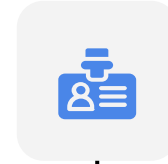
4. Source: <https://www.statista.com/outlook/dmo/eservices/dating-services/online-dating/worldwide>

5. Source: <https://www.ftc.gov/business-guidance/blog/2024/02/facts-about-fraud-ftc-what-it-means-your-business>

6. Source: <https://www.statista.com/statistics/677466/digital-ad-fraud-cost/>

7. Source: <https://www.ftc.gov/business-guidance/blog/2024/02/love-stinks-when-scammer-involved>

# User Research Findings



## Internet Platform Users

92% of users would use this service if it is **free**

82% of users would use this if it is **built into online media platforms**

*(User Survey)*

## Internet Company Employees

AI-driven scams / misinformation is a **big concern**

**2024 elections** are a high priority

*(Employee Interviews)*



# Web App MVP

## High Level Process



User Uploads Image



Face Detection



*for each face detected*

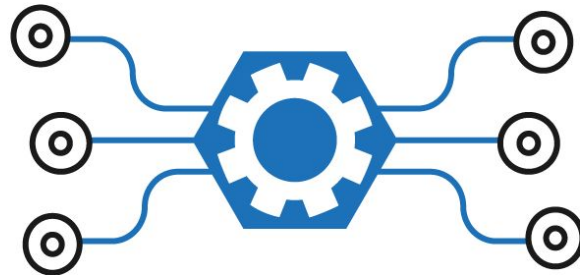


Result Returned to User

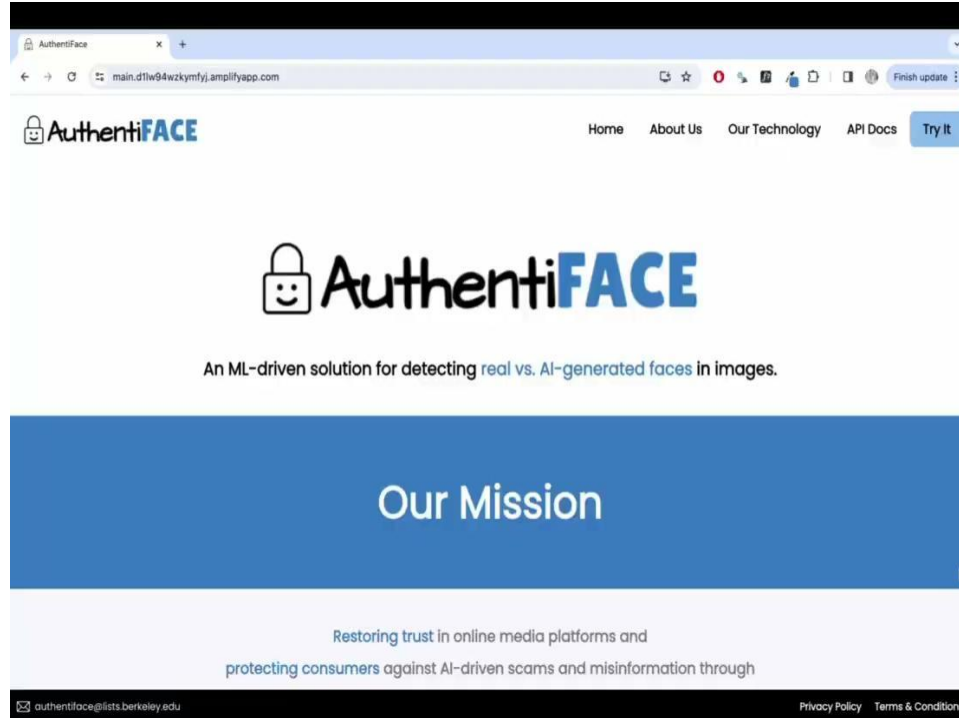
# AuthentiFace API MVP

Internet users would use our product if was **free** and built into **online media platforms**

- Can process millions of images per day
- Uses API keys for authentication and usage tracking
- Accepts image files, image URLs, and base64 encoded images
- Can batch process images




# MVP Demo



The screenshot shows a web browser window with the URL `main.d1lw94wzkyfj.amplifyapp.com`. The page features the AuthentiFACE logo at the top left and a navigation menu with links for Home, About Us, Our Technology, API Docs, and a blue 'Try It' button. The main content area displays the AuthentiFACE logo and the tagline: 'An ML-driven solution for detecting real vs. AI-generated faces in images.' Below this is a blue banner with the text 'Our Mission'. At the bottom, a light blue section contains the text: 'Restoring trust in online media platforms and protecting consumers against AI-driven scams and misinformation through'. The footer includes the email `authenti@lists.berkeley.edu` and links for 'Privacy Policy' and 'Terms & Conditions'.

AuthentiFACE

Home About Us Our Technology API Docs Try It

AuthentiFACE

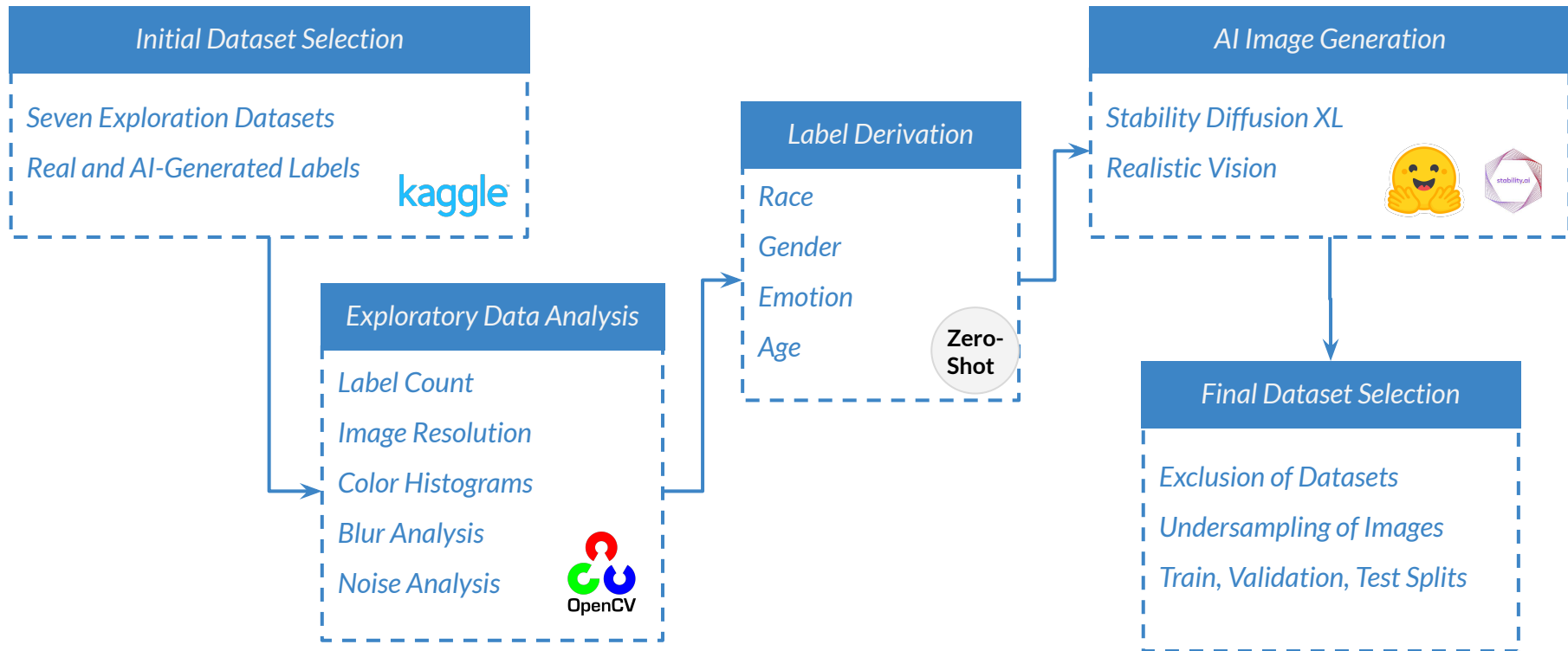
An ML-driven solution for detecting real vs. AI-generated faces in images.

## Our Mission

Restoring trust in online media platforms and  
protecting consumers against AI-driven scams and misinformation through

[authenti@lists.berkeley.edu](mailto:authenti@lists.berkeley.edu) [Privacy Policy](#) [Terms & Conditions](#)

# EDA, ML, and AI in Dataset Preparation



# Dataset Selections after EDA

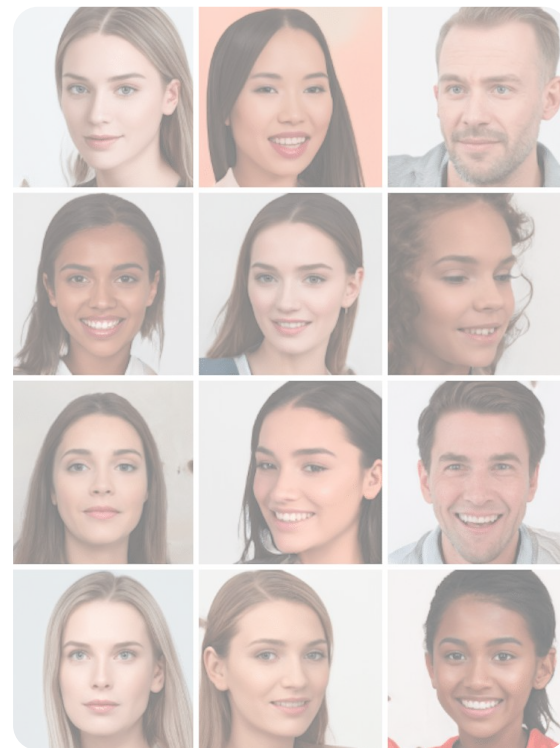
273K

Real Images

87K

AI-Generated Images

- [Face Dataset Of People That Don't Exist - Kaggle](#)
- [140k Real and Fake Faces - Kaggle](#)
- [Fake-Vs-Real-Faces \(Hard\) - Kaggle](#)
- [Person Face Dataset \(thispersondoesnotexist\) - Kaggle](#)
- [Large-scale CelebFaces Attributes \(CelebA\) Dataset - Multimedia Laboratory, The Chinese University of Hong Kong](#)
- ~~[Generated Faces - V7 Labs \(access issues\)](#)~~
- ~~[DigiFace 1M: 1 Million Digital Face Images for Face Recognition - Microsoft \(synthesized using untraditional methodology\)](#)~~



# Undersampling & AI Resolve Dataset Imbalance

273K

Real Images



## Undersample

Dataset of 'Real' Celebrity Faces

200K Faces of 10K Celebrities

Undersampled 10K images

87K

AI Generated Images



## Synthesize

AI-Generated Faces

Stable Diffusion and Realistic Vision

to generate 23K images



X

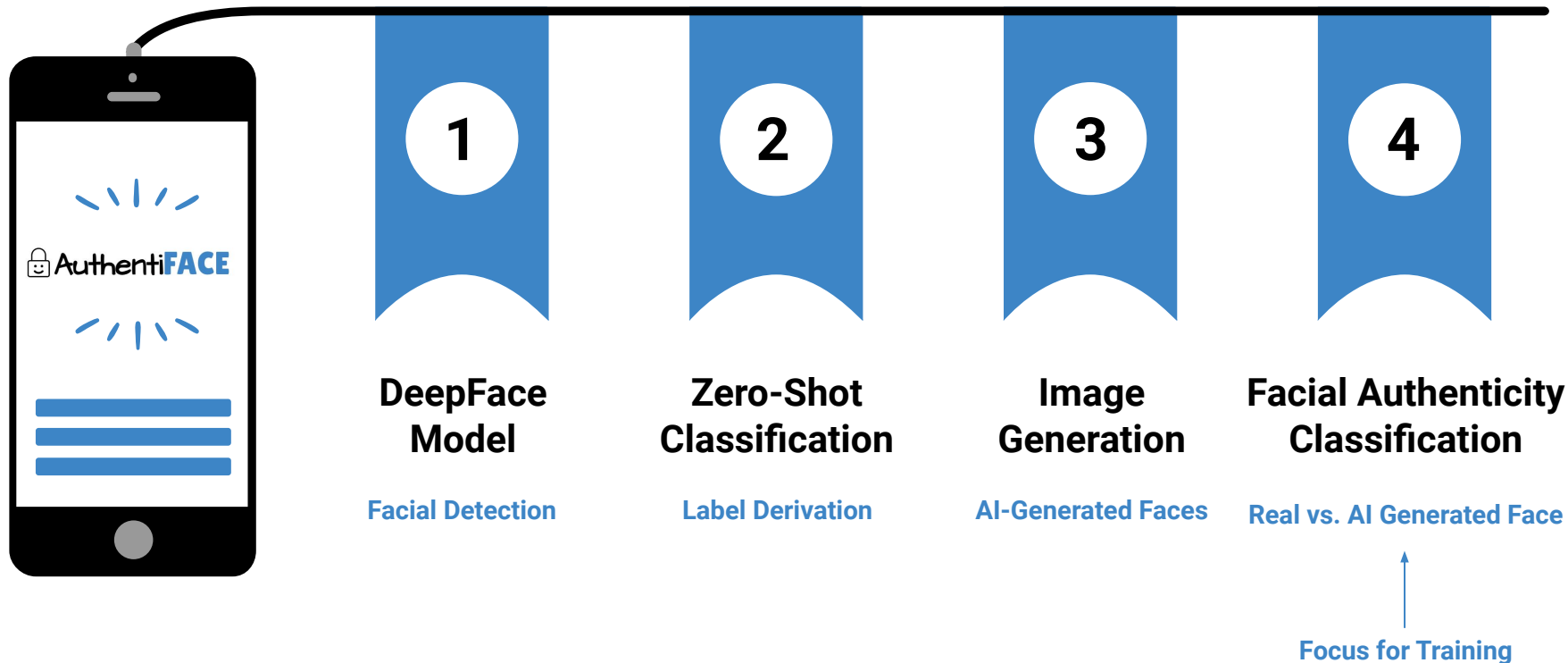


# Resulting Demographics after Label Derivation on our Training Dataset

*Undersampled to around ~100K for model training efficiency*

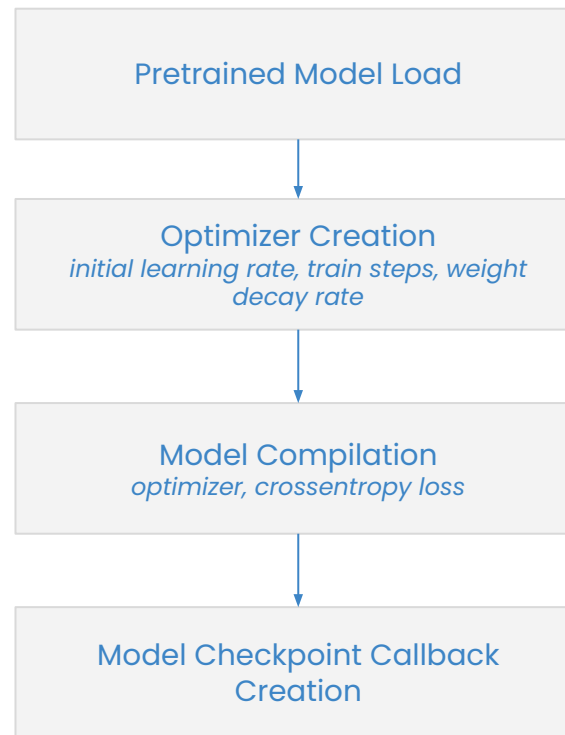
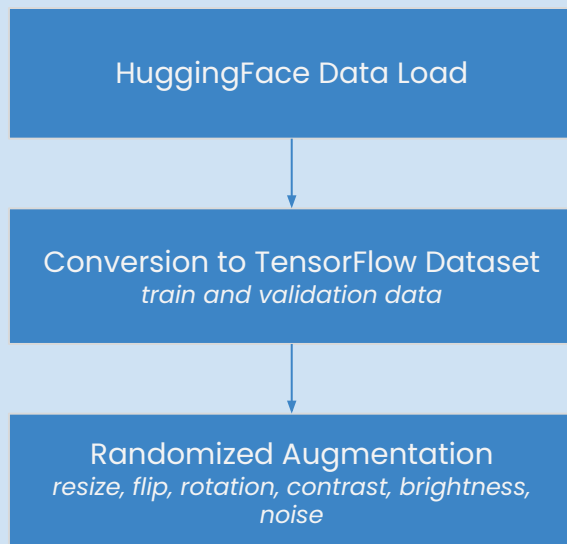
Image Label	Real (48,337 images)	AI-Generated (48,894 images)
<b>Derived Genders</b>	Male: 44.98% , <b>Female: 55.02%</b>	Male: 43.77% , <b>Female: 56.23%</b>
<b>Derived Races</b>	Latino Hispanic: 7628 (15.78%) <b>White: 21776 (45.05%)</b> Asian: 5491 (11.36%) Indian: 3955 (8.18%) Black: 5196 (10.75%) Middle Eastern: 4291 (8.88%)	Latino Hispanic: 10014 (20.48%) <b>White: 25444 (52.04%)</b> Asian: 5529 (11.31%) Indian: 2128 (4.35%) Black: 3522 (7.20%) Middle Eastern: 2257 (4.62%)
<b>Derived Emotion</b>	Happy: 2373 (4.91%) Angry: 123 (0.25%) Disgust: 1019 (2.11%) Surprise: 1121 (2.32%) Fear: 139 (0.29%) <b>Neutral: 43368 (89.72%)</b> Sad: 194 (0.40%)	Happy: 1014 (2.07%) Disgust: 68 (0.14%) Angry: 4 (0.01%) Fear: 26 (0.05%) Surprise: 256 (0.52%) <b>Neutral: 47515 (97.18%)</b> Sad: 11 (0.01%)
<b>Derived Age</b>	0-10: 0 (0.00%) 11-20: 191 (0.40%) <b>21-40: 41006 (84.83%)</b> 41-65: 7130 (14.75%) 66+: 10 (0.02%)	0-10: 0 (0.00%) 11-20: 512 (1.05%) <b>21-40: 42622 (87.17%)</b> 41-65: 5755 (11.77%) 66+: 5 (0.01%)

# Machine Learning in AuthentiFACE

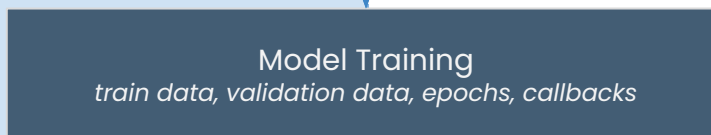




# ML Training Pipeline



Data



Model

## ResNet-50 Selected as Facial Authenticity Classification Model

Model	Epochs	Validation Accuracy	Training Time	GPU	Inference Time (ms)
ViT-base	16	91%	29 min/epoch	A10	120
				T4	184
ViT-large	16	93%	52 min/epoch	A10	229
				T4	376
Dino-vitb16	16	90%	29 min/epoch	A10	120
				T4	184
Swin-base	10	<b>99.5%</b>	1 hr 11 min/epoch	A10	570
				T4	805
<b>ResNet-50</b>	<b>16</b>	<b>99%</b>	<b>20 min/epoch</b>	<b>A10</b>	<b>66</b>
				<b>T4</b>	<b>102</b>

Unexpectedly, ResNet-50 had the second highest validation accuracy and the highest inference speed

# Training ResNet-50 on our Full Training Dataset

~196K real and AI-generated images from Kaggle & University of Hong Kong

Image Label	Real (85,015 images)	AI-Generated (110,693 images)
<b>Derived Genders</b>	Male: 45.75% , <b>Female: 54.25%</b>	Male: 44.22% , <b>Female: 55.78%</b>
<b>Derived Races</b>	Latino Hispanic: 8245 (9.70%) <b>White: 45888 (53.98%)</b> Asian: 12725 (14.97%) Indian: 4993 (5.87%) Black: 6139 (7.22%) Middle Eastern: 7025 (8.26%)	Latino Hispanic: 18260 (16.50%) <b>White: 60458 (54.62%)</b> Asian: 16243 (14.67%) Indian: 3259 (2.94%) Black: 5812 (5.25%) Middle Eastern: 6661 (6.02%)
<b>Derived Emotion</b>	Happy: 4280 (5.03%) Angry: 253 (0.30%) Disgust: 3191 (3.75%) Surprise: 2038 (2.40%) Fear: 446 (0.52%) <b>Neutral: 74222 (87.30%)</b> Sad: 585 (0.30%)	Happy: 2182 (1.97%) Disgust: 332 (0.30%) Angry: 16 (0.01%) Fear: 117 (0.11%) Surprise: 569 (0.51%) <b>Neutral: 107420 (97.04%)</b> Sad: 57 (0.05%)
<b>Derived Age</b>	0-10: 0 (0.00%) 11-20: 323 (0.38%) <b>21-40: 71709 (84.35%)</b> 41-65: 12962 (15.25%) 66+: 21 (0.02%)	0-10: 0 (0.00%) 11-20: 1108 (1.00%) <b>21-40: 95959 (86.69%)</b> 41-65: 13615 (12.30%) 66+: 11 (0.01%)

# High Accuracy, Precision, and Recall

## Test Set Accuracy

99.3%

*20k images*

Class	Precision	Recall	F1
Real	98.92%	99.4%	99.2%
Fake	99.57%	99.2%	99.4%

# Explainability through Integrated Gradients

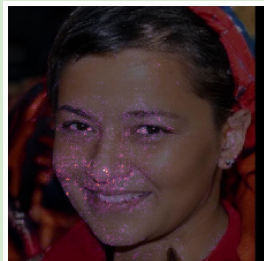
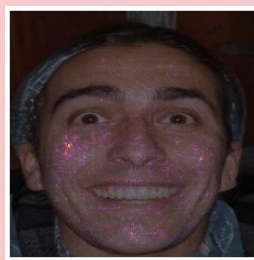
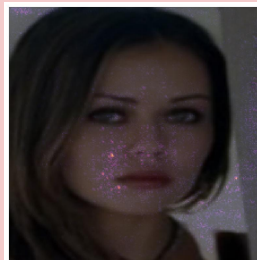
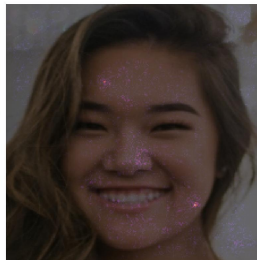
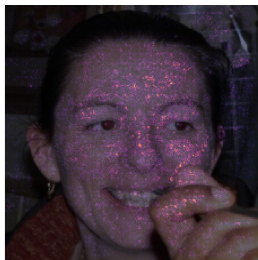
True Positive

Rate = 0.9942

Labeled as Real

False Positive

Rate = 0.0043



True Negative

Rate = 0.992

Labeled as Fake

False Negative

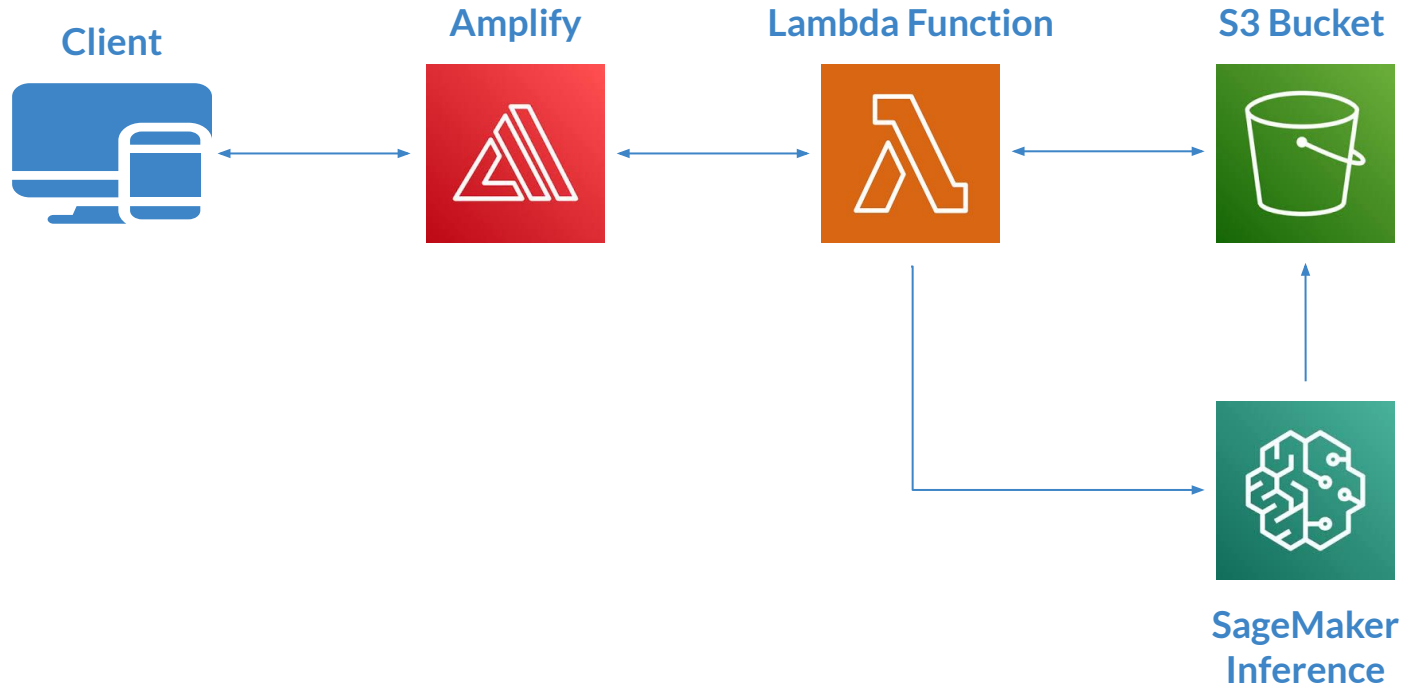
Rate = 0.0058

# Interpretation: Higher In-Domain Performance

- Our model excels on real and fake faces similar to our dataset but underperforms on unfamiliar styles
- Recent publications, such as "Finding AI-Generated Faces in the Wild,"<sup>8</sup> highlight similar challenges
- Ethical concerns limit access to diverse, high-quality real face datasets
- Existing datasets lack professional headshots, leading to classification issues

Model	In-Domain Recall	Out-of-Domain Recall
AuthentiFace	99.29%	86.67%
Finding AI-Generated Faces in the Wild <sup>8</sup>	98.0%	84.5%

# Polling Architecture through AWS



# Overcoming Technical Challenges

## Dataset

1

A diverse, balanced dataset that is representative of user images

## Architecture

2

A robust and secure architecture that supports bulk image classification

## Scalability

3

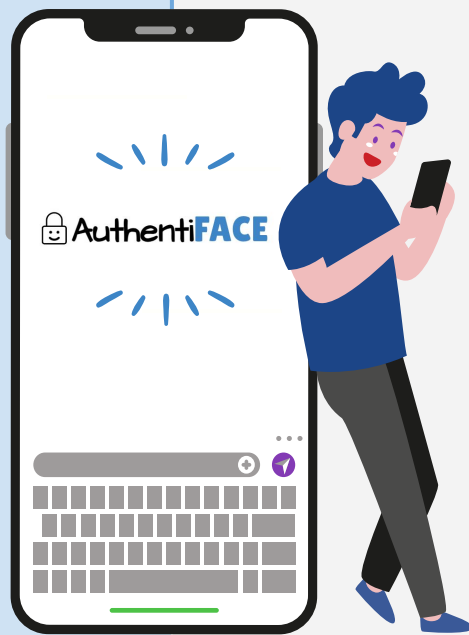
An integrated, explainable solution that works out of domain





# Overall User AuthentiFace Rating = 4.5

- Model performs lower on **out-of-domain** images
- Some images that should have produced **errors were classified as real**
- Website **easy to navigate** and professional-looking

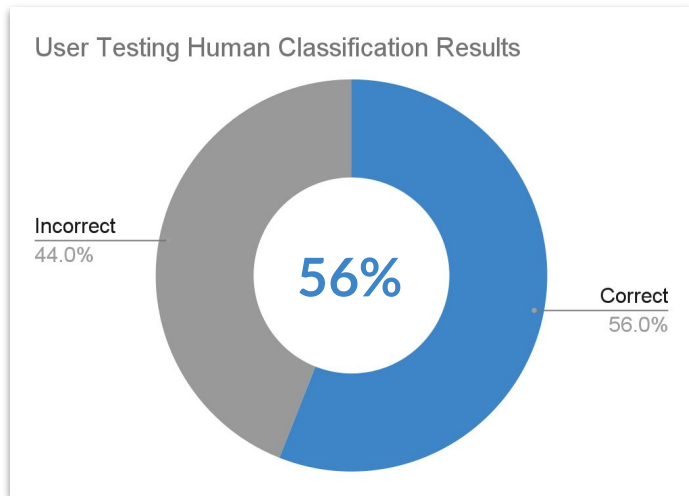


- Interest in **drag-and-drop** photo upload
- Interest in reasons **why images are classified** as real vs. fake
- Interest in ability to **upload multiple images at once**

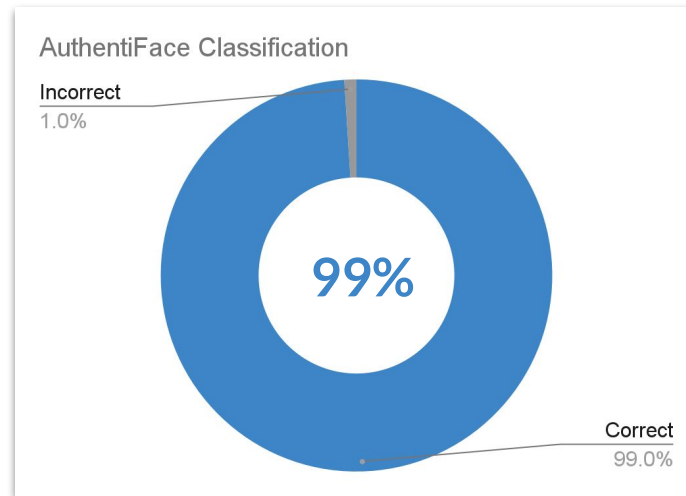
Observations

Opportunities

# AuthentiFace **Outperforms** Human Classification



VS.



AuthentiFace accuracy significantly outperforms manual human classification.

# What's Next?

## Diversify Data

to include race, age, disability,  
and other minority classes

1

## Collect Images

of real faces with varying  
image quality

2

## Generate Images

of AI faces using additional  
models and prompts

3



4

## Enhance UI

by implementing user  
testing feedback

5

## Provide Explainability

by including integrated  
gradients in user results

**Dataset**

**User Experience**



Restoring trust in online media platforms and protecting consumers against AI-driven scams and misinformation through facial authenticity detection.