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

Ghiwa Lamah  Product Manager
Nancy Ding  User Experience
Melissa Olivera  Data Expert
Jacob Petrisko  ML Expert
Queen Tran  Architecture Lead

EDA, Explainability  Market & User Research  User Testing  Model Evaluation  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

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
- $1.4 Billion reported losses due to fraud on social media in the US in 2023

**Digital Advertising**
- $740.3 Billion projected ad spending on digital advertising by 2024
- $84 Billion estimated cost of ad fraud worldwide in 2023

**Dating Apps**
- $3.12 Billion projected revenue of online dating market by 2024
- $1.1 Billion reported losses due to romance scams in the US in 2023

---

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

Crop Image to Face

Image Quality Checks

Facial Authenticity Prediction

for each face detected

Face is REAL

Result Returned to User
AuthentiFace API MVP

Internet users would use our product if it 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

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

AuthentiFACE
EDA, ML, and AI in Dataset Preparation

Initial Dataset Selection
- Seven Exploration Datasets
- Real and AI-Generated Labels

Exploratory Data Analysis
- Label Count
- Image Resolution
- Color Histograms
- Blur Analysis
- Noise Analysis

Label Derivation
- Race
- Gender
- Emotion
- Age

AI Image Generation
- Stability Diffusion XL
- Realistic Vision

Final Dataset Selection
- Exclusion of Datasets
- Undersampling of Images
- Train, Validation, Test Splits
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
Resulting Demographics after Label Derivation on our Training Dataset
*Undersampled* to around ~100K for model training efficiency

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

1. DeepFace Model
   - Facial Detection

2. Zero-Shot Classification
   - Label Derivation

3. Image Generation
   - AI-Generated Faces

4. Facial Authenticity Classification
   - Real vs. AI Generated Face

Focus for Training
ML Training Pipeline

1. HuggingFace Data Load
2. Conversion to TensorFlow Dataset
   - train and validation data
3. Randomized Augmentation
   - resize, flip, rotation, contrast, brightness, noise
4. Pretrained Model Load
5. Optimizer Creation
   - initial learning rate, train steps, weight decay rate
6. Model Compilation
   - optimizer, crossentropy loss
7. Model Checkpoint Callback Creation
8. Model Training
   - train data, validation data, epochs, callbacks
<table>
<thead>
<tr>
<th>Model</th>
<th>Epochs</th>
<th>Validation Accuracy</th>
<th>Training Time</th>
<th>GPU</th>
<th>Inference Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-base</td>
<td>16</td>
<td>91%</td>
<td>29 min/epoch</td>
<td>A10</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T4</td>
<td>184</td>
</tr>
<tr>
<td>ViT-large</td>
<td>16</td>
<td>93%</td>
<td>52 min/epoch</td>
<td>A10</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T4</td>
<td>376</td>
</tr>
<tr>
<td>Dino-vitb16</td>
<td>16</td>
<td>90%</td>
<td>29 min/epoch</td>
<td>A10</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T4</td>
<td>184</td>
</tr>
<tr>
<td>Swin-base</td>
<td>10</td>
<td>99.5%</td>
<td>1 hr 11 min/epoch</td>
<td>A10</td>
<td>570</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T4</td>
<td>805</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>16</td>
<td>99%</td>
<td>20 min/epoch</td>
<td>A10</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T4</td>
<td>102</td>
</tr>
</tbody>
</table>

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

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

Test Set Accuracy

99.3%

20k images

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>98.92%</td>
<td>99.4%</td>
<td>99.2%</td>
</tr>
<tr>
<td>Fake</td>
<td>99.57%</td>
<td>99.2%</td>
<td>99.4%</td>
</tr>
</tbody>
</table>
Explainability through Integrated Gradients

- **True Positive Rate**: 0.9942
- **True Negative Rate**: 0.992
- **False Positive Rate**: 0.0043
- **False Negative Rate**: 0.0058

**Labeled as Real**

- Real images labeled as Real
- Real images labeled as Fake

**Labeled as Fake**

- Fake images labeled as Real
- Fake images labeled as Fake
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," highlight similar challenges.
- Ethical concerns limit access to diverse, high-quality real face datasets.
- Existing datasets lack professional headshots, leading to classification issues.

<table>
<thead>
<tr>
<th>Model</th>
<th>In-Domain Recall</th>
<th>Out-of-Domain Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>AuthentiFace</td>
<td>99.29%</td>
<td>86.67%</td>
</tr>
<tr>
<td>Finding AI-Generated Faces in the Wild</td>
<td>98.0%</td>
<td>84.5%</td>
</tr>
</tbody>
</table>

Polling Architecture through AWS

- Client
- Amplify
- Lambda Function
- S3 Bucket
- SageMaker Inference
Overcoming Technical Challenges

**Dataset**
A diverse, balanced dataset that is representative of user images

**Architecture**
A robust and secure architecture that supports bulk image classification

**Scalability**
An integrated, explainable solution that works out of domain
Overall User AuthentiFace Rating = 4.5

Observations:
- 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

Opportunities:
- 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
AuthentiFace **Outperforms** Human Classification

**User Testing Human Classification Results**
- Incorrect: 44.0%
- Correct: 56.0%

**AuthentiFace Classification**
- Incorrect: 1.0%
- Correct: 99.0%

AuthentiFace accuracy significantly outperforms manual human classification.
What’s Next?

1. **Diversify Data**
   to include race, age, disability, and other minority classes

2. **Collect Images**
   of real faces with varying image quality

3. **Generate Images**
   of AI faces using additional models and prompts

4. **Enhance UI**
   by implementing user testing feedback

5. **Provide Explainability**
   by including integrated gradients in user results

**Dataset**

**User Experience**
AuthentiFACE

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