# RUFake: Online Tool for Detecting AI Generated or Altered Photos

**Team members**

[DJ Fajana](https://www.ischool.berkeley.edu/people/dj-fajana)

[Qinchen Jiang](https://www.ischool.berkeley.edu/people/qinchen-jiang)

[Jasmine Lau](https://www.ischool.berkeley.edu/people/jasmine-lau)

[Sanjiv Narayan](https://www.ischool.berkeley.edu/people/sanjiv-narayan)

[Tamar Brand-Perez](https://www.ischool.berkeley.edu/people/tamar-brand-perez)

## Teaser text:

RUFake detects deepfakes to protect you from online scams—fast, private, accurate, and built with ethics and zero data storage.

## Teaser Image

## 

## Elevator Pitch

RUFake protects you from online scams by identifying a deepfake photo from a real one - rapidly, accurately and with your privacy and ethics built in. The Economist recently called online scams the “largest illicit industry in the world,” valued above $500 billion (1). Online scams including romance scams and business scams have cost individuals their life savings and left them emotionally devastated. RUFake has state-of-the-art AI that we developed using unbiased analytics from archives of millions of images. We do not store any data, we do not take advertising money, and nobody can ever hack us for your data - as we don’t have it. Use RUFake - because we don’t want you to be the next victim

## Problem & Motivation

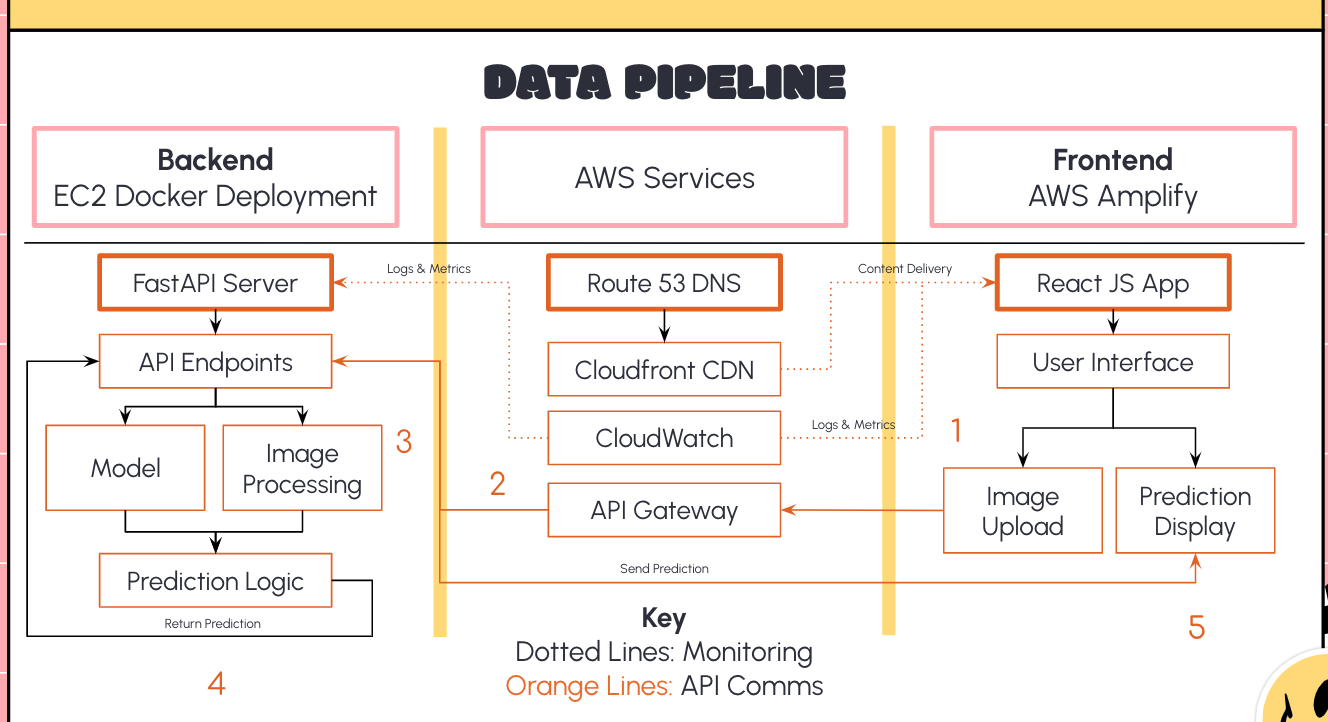
Online scamming is a $500B global challenge that can cause financial ruin - and emotional distress. 53% of men & 47% of women are victims of a romance scam. Users require a trustworthy safe solution to detect scamming and protect them.

## RUFake

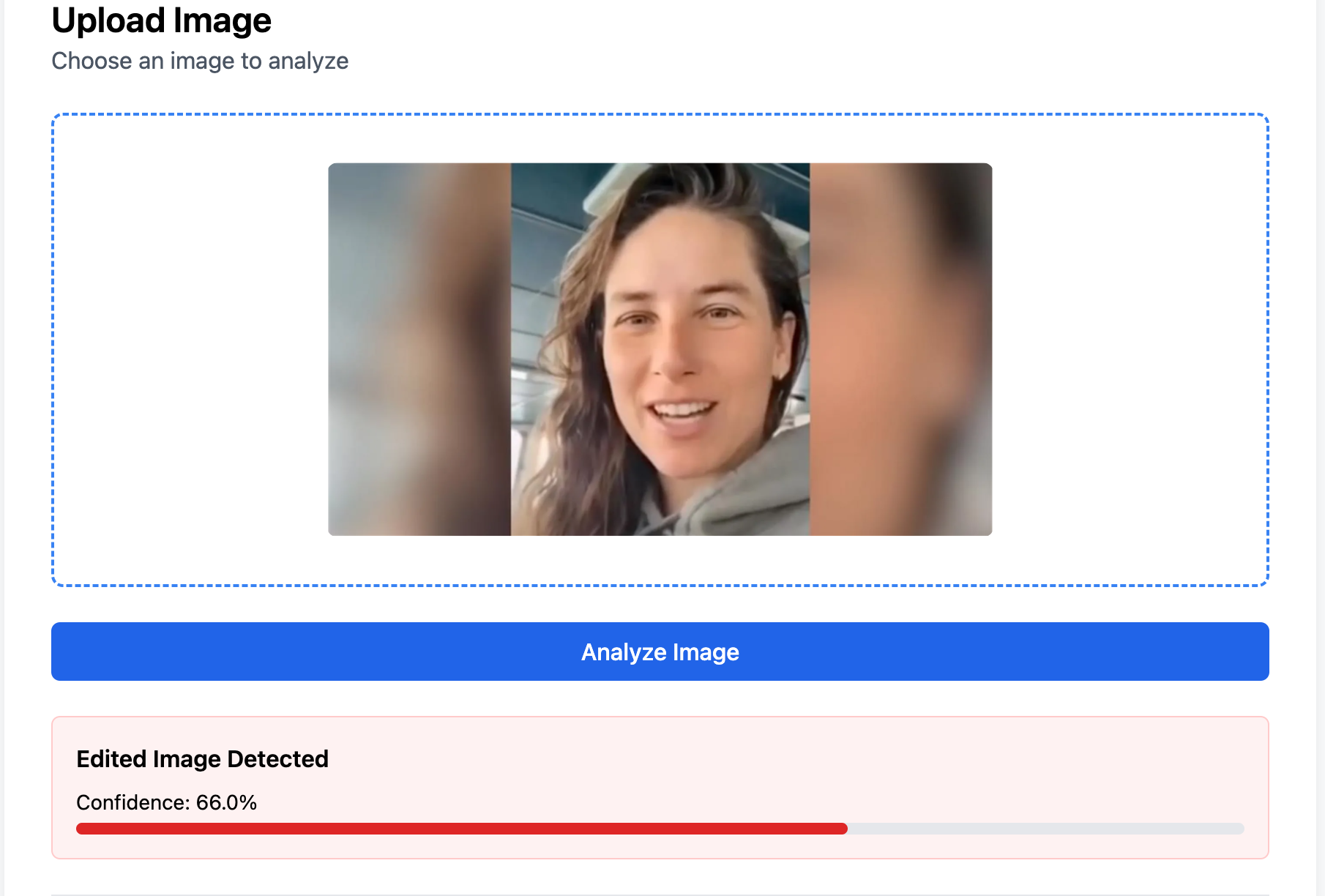
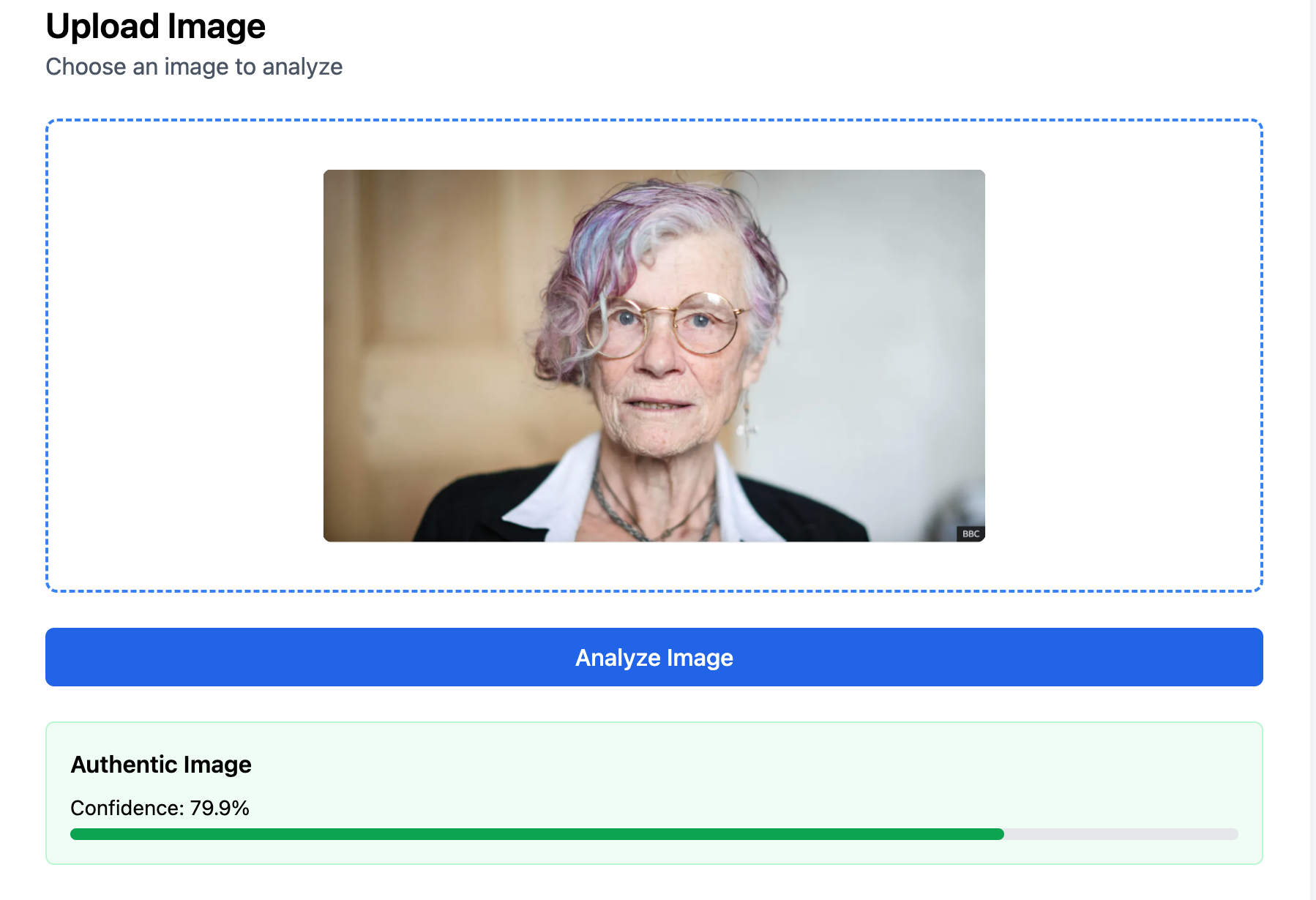
Addresses several unmet needs by providing an online tool with >85% accuracy to detect real vs fake photos of individuals you interact with. Our primary use case is online dating; our secondary use case is online ID checking. We prioritize privacy - nobody will know you are checking. We prioritize security - we don’t store your profile or photos. Our AI was built from millions of photos in broad, diverse populations.

RUFake is a web application where users can upload images and receive a prediction & confidence of whether an image is AI generated. Since it is web based, users can screenshot images from dating apps and upload them directly to the site and get confirmation in less than 3 seconds.

RUFake is a ReactJS web application deployed by AWS Amplify, hosted with AWS Route 53 services on rufakeapp.com. Once an image is uploaded on the frontend, it is sent via an api gateway to our EC2 instance, where a FastAPI server has our ViT model loaded. From there FastAPI validates the image input, and returns a prediction back through the API gateway to our end user in less than 3 seconds. The outputs are confidence scores as well as a determination of whether or not the image is fake.



Below are examples of the prediction display and outputs. Users will see green if it is likely that an image is authentic, and red if it is likely AI generated. The degree of confidence is designated by a status bar below the prediction. A prediction cannot below 50%, similar to a softmax, which is our threshold, and extreme confidence typically arrives around 80%.



## Privacy and Ethics (By Design)

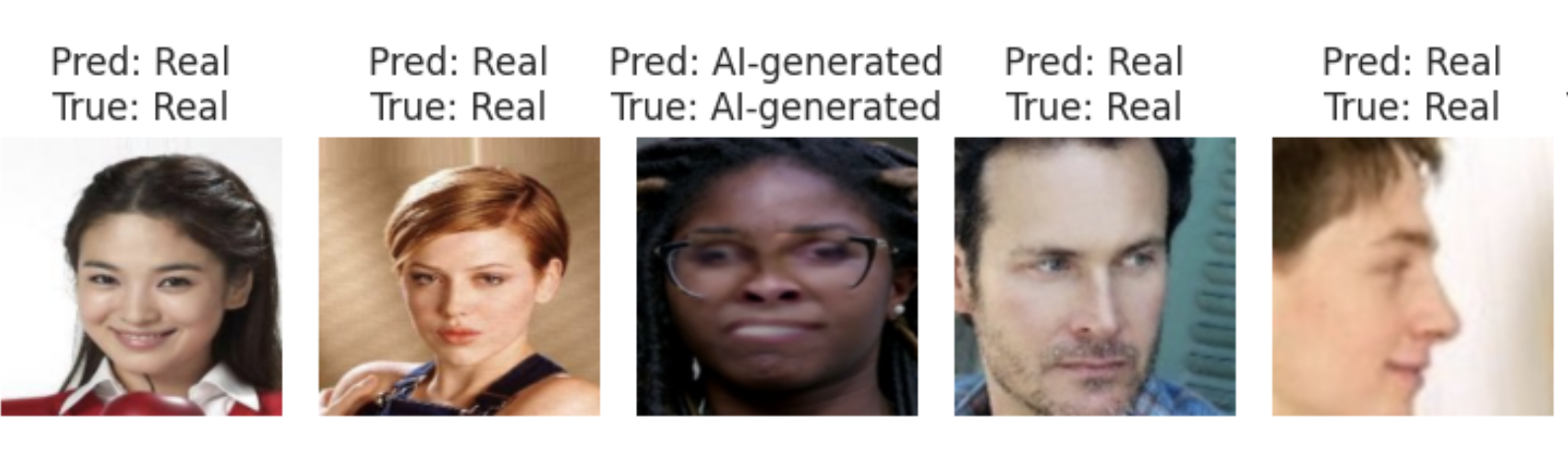
* As described by Nissenbaum (2), this project requires privacy by design, for which strategies are well outlined (3).
* The purpose of RUFake is to distinguish real from AI generated images. While this technology offers tools for information verification, it could cause emotional distress, financial disruption or embarrassment in potentially vulnerable populations (4).
* Privacy and ethics concerns result from:
  + Vulnerability of users - Revealing to a vulnerable user that a photo is manipulated;
  + Risk of incorrect determinations - A need to mitigate the legal, financial and emotional impact of incorrect determinations (app errors). Mitigation is difficult, even using detailed explanations (6).
  + Bias in detection models - Risks of biased analysis, that preferentially tag photos by certain individuals or locations as fake. Risks can be legal, financial or psychological, and impact the user, RUFake and data owners. Mitigation must avoid perpetuating structural biases at several levels (7). These risks exist even if bias was introduced inadvertently through class imbalances in our training data.
  + Data security and consent - Must keep the photo, app determination and user’s data secure. We will provide disclaimers that it is difficult to ascertain the source of photos, particularly if fake. This will mitigate the risks to RUFake of storing or analyzing sensitive or even illegal data.
  + Disclosures and implications of identity - Identity, membership or attribute disclosures could have secondary or implied meaning as outlined in a recent publication by members of the RUFake team (8).
* Our Solution thus minimized legal exposure (5) and emotional distress for the user, the subject of the photo and the owner of the photo.
  + Disclaimers for errors (inaccurate authenticity determination), including the limitations of training data, limited languages or dataset sizes.
  + No data storage. This avoids the need to prevent data leakage.
  + Reduce risks of aggregation, as outlined for specific data types in our recent paper8.
  + How long should data security last? This is open ended but may not be relevant as we will not store data. Similarly, data review, editing and deletion is not relevant.
  + Diverse training data and fairness-aware learning techniques are employed.
  + We included stakeholder input from a diverse group of people.
  + Educational material accompanies the app to guide ethical usage and interpretation of results.  
    Future revisions could implement a mitigation plan for inaccurate authenticity determination. This may require image data storage for analysis, which we currently avoid.

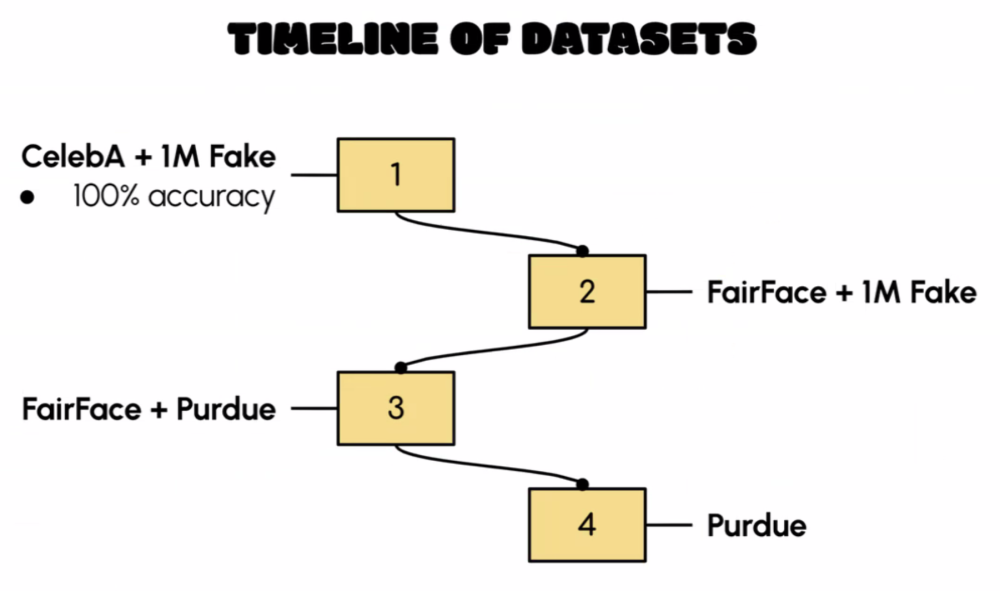
## Data Sources & Data Science Approach

Our dataset selection process underwent several stages of research and development. Each stage was shaped by both our technical goals and the limitations we encountered. An early objective of ours was to train a model that could classify real and fake images with a less than 10% difference in accuracy between gender (male/female) and race (white/non-white) given that prior studies have highlighted an overrepresentation of white-male images in similar models.

To begin our search for suitable datasets, we initially selected CelebA (9) as our real image dataset due to its size and widespread use in academic research. For fake images, we used the 1 Million AI Generated Faces 128x128 (10) dataset, chosen for being large in scale and equally accessible. However, neither CelebA nor the 1 Million images included explicit annotations for gender or race, which are necessary to make our balanced classifications. To address this, we used DeepFace (11), an open-sourced facial analysis framework that bundles several state-of-the-art models (VGG-Face, FaceNet, ArcFace, etc.) to extract our desired attributes (gender and race).  
  
After processing both datasets through DeepFace, we constructed a dataset that is balanced for gender and race. This processing and feature engineering resulted in a dataset with ~80,000 images, which we then proceeded to train and evaluate our model on. Surprisingly, the model that had been trained on this processed dataset achieved 100% classification accuracy, which raised suspicions of overfitting or potential data leakage.  
  
Taking a closer look at our datasets, we discovered that CelebA contains over 200,000 images, but only ~10,000 unique faces, making it likely that the same faces appeared in our training, validation, and test splits. Following guidance from our professors and unsuccessful attempts to resolve this issue, we pivoted to other datasets to cover the real faces part of our dataset.  
  
We turned to FairFace, like CelebA, is a robust dataset, but was designed to provide balanced representations across race and gender (12). FairFace aligns with our initial goal of achieving a model that provides near equal classification accuracy between gender and race. Coming pre-annotated with these attributes, we used FairFace for our real images and selected a matching number of images from the 1 Million to maintain balance between real and fake images.  
  
However, the challenge of our model producing a 100% accuracy persisted, which led us to dig further into the causes of this issue. We found that the 1 Million images were all generated solely with NVIDIA’s StyleGAN (10). Each generation method produces distinct ‘fingerprints’ (13) that can make classification easy for our model. Since real images do not have these fingerprints, they become immediately distinguishable.   
  
Since the previous iterations of our datasets proved to be lacking, we decided to again, pivot datasets and use FairFace alongside Purdue’s AI-Face (We received approval for using Purdue’s dataset mid semester). Purdue’s dataset offered several advantages compared to our previous 1 Million dataset. This dataset from Purdue (14) has both real and fake images, where the fake images were generated using 37 distinct methods and has annotations. The main challenge with this dataset is that it is very large (>300GB) and not easily accessible which led to significant slowdown in our process.   
  
Nonetheless, the processing of ~40,000 real images from FairFace and ~40,000 fake images from Purdue still reached a 100% classification accuracy, so we used the Purdue dataset alone, which contains both real and fake images with the same annotation scheme.   
  
  
Through the research of these dataset iterations, we also identified future work. We are interested in minimizing the impact of the identifiers left by generation methods of AI generated images. Two methods we would like to look further into are: removing the identifiers from fake images using tools such as GANprintR (15), or injecting similar fingerprints into real images (16). We have successfully experimented with fingerprint injection with FairFace and were able to use this data in our model, but we ultimately deemed it outside the scope of our project since we had already met our primary goals we had set for our project.

* Figure 1 below shows examples from our data repository of real and AI generated images.





In summary, our final dataset contains:

* 39,436 real || 27,914 fake
* 33,621 female || 33,729 male
* 34,171 white || 33,179 non-white
* Age distribution: young 43,777 || middle-aged 9,665 || senior 1,262 || other 12,646

Images of varying sizes were all changed to 224 x 224 to fit the ViT\_base\_patch16-224 model.

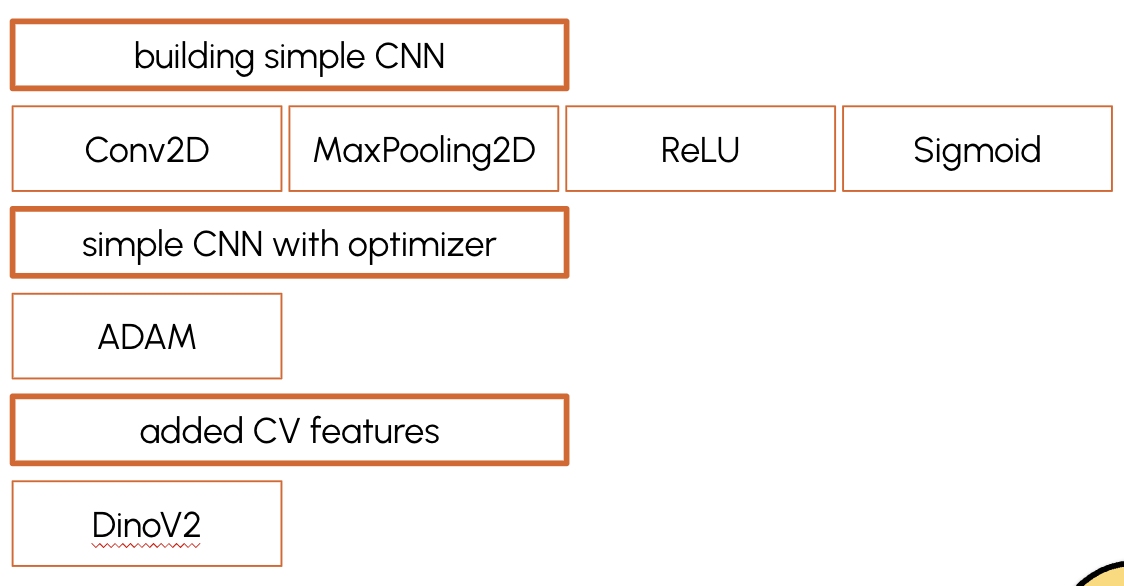
* These images were randomly chosen from the dataset and were balanced for gender and ethnicity to minimize bias.

## Model Design and Development

**Model Design**

* We began by selecting a baseline model and tested several architectures—VGG16, ResNet50, DenseNet121, MobileNet, InceptionV3, and a simple CNN—using the initial set of 2,000 labeled images (real/fake). Evaluations with SGD and ADAM optimizers showed ResNet50 performed slightly better, but most models achieved similar accuracy (50–55%). We chose the simple CNN with ADAM due to its lower parameter count, faster deployment, and suitability for rapid prototyping, particularly for potential mobile integration.
* **Model structure.** We had a total of 7 layers and processes, and added DinoV2 for image featuring:
* **Layers and process.** Cov2D, MaxPooling2D, Conv2D, MaxPooling2D, Flatten (not a layer), Relu "Activation Function", Sigmoid "Activation Function" (Output Layer).

Our CNN model architecture is shown below.



* **Featurization.** Several computer vision approaches we tested include: complex computer vision features, DinoV2, and some simple computer vision features: Canny edge detection, HOG, Fourier Transform, ELA, edge orientation distribution, color\_distribution, dominant\_colors. Only Dinov2 improved the model accuracy by 6% so we kept it in our CNN model.

******

* **Change of training dataset.** From here, we started to train the CNN model with the 80K photos our team finished processing with annotations and new labels.

## Final model

## ViT Base Patch16 224

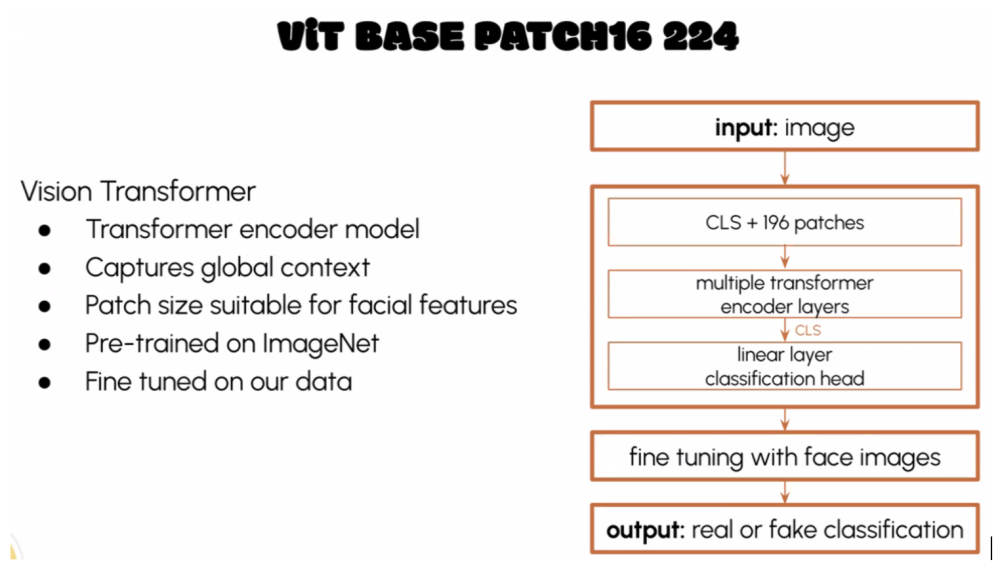
We chose to use the vision transformer in our application. It’s a transformer encoder model.

It treats the image as a sequence of patches making it excellent at capturing global context as well as subtle spatial patterns through attention. The size of the patches is suitable for capturing facial features.

Pretraining on image-net gives it general visual understanding of textures, edges and shapes,

and fine tuning it on our data allowed us to adapt the model to our domain.

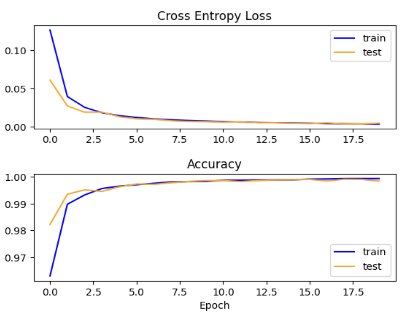
It also provides a good balance between performance and computational efficiency. It is not very small but it was fast enough to perform at 1 second per inquiry. The model learns high-dimensional hierarchical features. Some features it could pick up for the fake images are: inconsistent lighting, misaligned facial features, and blurred or overly smoothed textures.



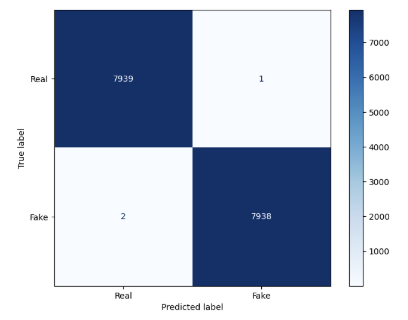
## Evaluation

**Evaluation of the CNN models:**

* **Training Process evaluation**. We looked at Loss and Accuracy at each epoch (total of 20 epochs). Our Loss curve and accuracy reached 100%.



* **Result Performance evaluation.** Precision, recall and confusion Matrix also showed more than expected results.

****

**Troubleshooting.** We focused on two key areas to achieve optimal training results: model logic and implementation, and dataset-related issues.

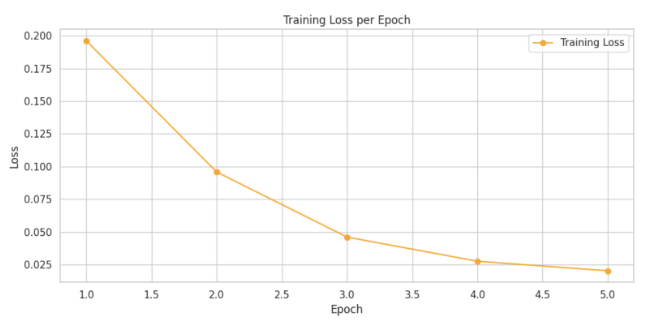
* **Model logic.** We tested with balanced photo sets ranging from 200 to 55,000 (all) images. The results showed a consistent increase in accuracy with larger training sets—an expected trend in model training—thereby ruling out model logic issues.
* **Data leakage.** Described in the dataset section.
* **StyleGAN fingerprint process.** Since all fake images were generated using StyleGAN, we hypothesized the model might be detecting hidden fingerprints or watermarks. To test this, we replaced the real images with those from the FairFace dataset and added a similar fingerprint. The model achieved 65% accuracy, supporting our hypothesis of bias from the original dataset. However, due to time constraints, we were unable to further refine the CNN model.

Evaluation of the final model:

Prior to fine tuning, the results roughly at 50%, demonstrated that the model was basically guessing if the image was real or AI generated. The hyperparameters we used to train our model were: Learning rate of 2e-5, small, suitable for fine tuning a pre trained model. It allowed for a slower and more precise process of adjusting the weights. We used Cross Entropy Loss that is suitable for a classification task.

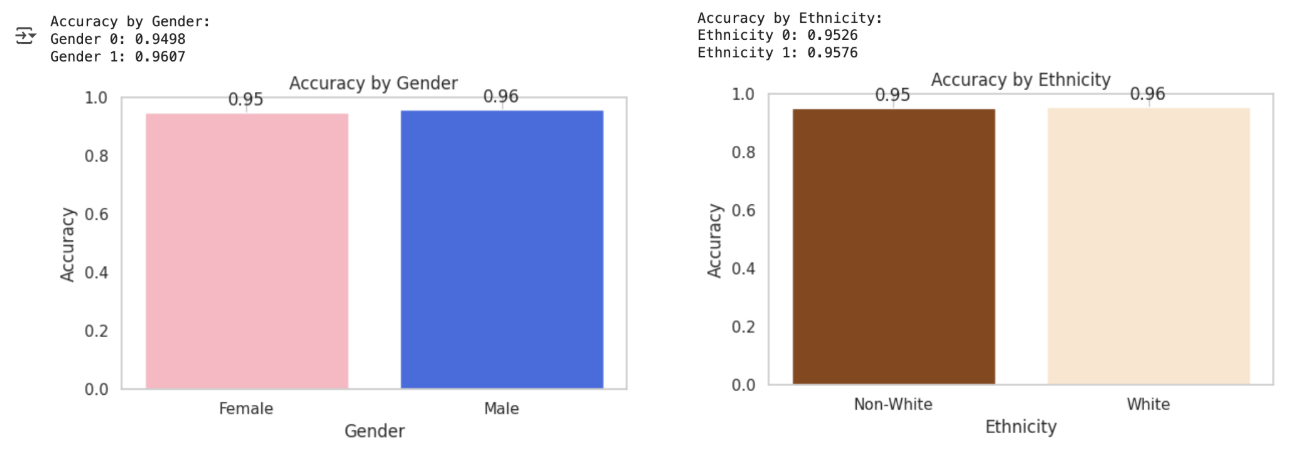
Batch size 32 and we used AdamW which is an improved version of the Adam optimizer, it handles weight decay by adding a penalty term to the loss function preventing overfitting.

Our final model accuracy is shown below:



The top image shows the model is gradually improving its performance with each epoch, during training. In the bottom image you can see the model learns well, and with a high test accuracy it generalizes well to the test set. After fine tuning with the new dataset, the model’s overall f1 score is around 96% showing high performance in distinguishing between real and fake images.

We tested the performance of the model across different demographic groups and found practically no difference between the groups. The model performed at around 95 or 96% for all groups.



Generalizability of this model beyond this dataset at this point seems good but not perfect. It was able to correctly recognize some real images and some fake images and cartoons not from the dataset, it was also able to correctly classify images that were professionally retouched and images with filters from apps like instagram and snapchat. However, when we asked it to recognize images created by Chatgpt which it was not trained on, it thought they were real.

Our model achieved strong results, outperforming existing models designed for detecting real vs. fake images—particularly in terms of fairness across demographic groups. On the test set, the model reached 95% accuracy. We set an initial goal of keeping performance differences between demographic groups below 10%, and we surpassed this with a difference of only ~1%, demonstrating excellent consistency and fairness.

In terms of speed, our target was to generate a prediction within 3 seconds. The final model exceeded expectations, delivering predictions in about 1 second per image. Additionally, we ensured that no images are stored, preserving user privacy and aligning with our ethical commitment to responsible AI.

## Key Learnings & Impact

* Finding the right dataset that would enable appropriate training for our model was difficult. We needed to consider not only the total number of images but also the number of people the images were of. Combining datasets in many versions did not work and the only large enough suitable dataset was one that we had to receive special permission to use which took time. It was also challenging to use this dataset due to its very large size.
* Balancing for gender and ethnicity in the dataset significantly improved the fair performance of our model compared to previous work done in the field.
* We demonstrated it is possible to distinguish between real and AI generated images in a robust and fair way.
* Privacy was surprisingly difficult to ensure.

## Conclusions

* We identified online scamming as a major societal and financial concern worldwide. To address this problem, we developed an app to separate real from AI-generated images.
* Our design mandate was to make the app simple-to-use, and to make privacy and ethics a design-feature.
* We tested a variety of de novo neural network architectures, feature extraction techniques and model sets to separate Real from altered/AI generated images of individuals.
* We deployed our best model on the web, which provided high performance.
* Our model received excellent feedback in early market testing.
* Our model stores no data from the user, photo provider nor photo subject, thus adhering to our privacy-by-design mandate.
* Our team worked well together, and according to plan.
* Future anticipated work will refine the model, with a view to potential commercialization.

## Acknowledgements

* Superb input from Prof Puja Vahabi and Prof Todd Holloway
* We are indebted to input from the Berkeley Computer Vision Faculty
* We acknowledge feedback from our peers who have shaped the project

## References

1. The Economist. (2025, February). *Online scamming*.

2. Nissenbaum, H. (2011, October). A Contextual Approach to Privacy Online. *Daedalus, 140*(4), 32–48.

3. Wright, D. (2012). The state of the art in privacy impact assessment. *Computer Law & Security Review, 28*(1), 54–61.

4. Allen, K. C., Davis, A., & Krishnamurti, T. (2023). Indirect Identification of Perinatal Psychosocial Risks from Natural Language. *IEEE Transactions on Affective Computing, 14*, 1506–1519. https://doi.org/10.1109/TAFFC.2021.3079282

5. Dankar, I. K., & El Emam, K. (2013). Practicing differential privacy in health care. *Transactions on Data Privacy, 6*, 35–67.

6. Caldwell, S., & Temmermans, F. (2024). Fake images: How much do they matter? Towards building quantitative modeling of misinformative images. Paper presented at *Applications of Digital Image Processing XLVII*, SPIE Proceedings, 13137(E).

7. Whong, C. (2014). *FOILing NYC’s taxi trip data*.<https://chriswhong.com/open-data/foil_nyc_taxi/>

8. Narayan, S. M., Kohli, N., & Martin, M. M. (2025). Addressing contemporary threats in anonymised healthcare data using privacy engineering. NPJ digital medicine, 8(1), 145. https://doi.org/10.1038/s41746-025-01520-6

9. S. Yang, P. Luo, C. -C. Loy and X. Tang, "From Facial Parts Responses to Face Detection: A Deep Learning Approach," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3676-3684, doi: 10.1109/ICCV.2015.419

10. Dullaz. (2021). 1M AI generated faces 128x128. Retrieved February 2025 from https://www.kaggle.com/datasets/dullaz/1m-ai-generated-faces-128x128

11. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., & Wu, B. (2023). *DeepfakeBench: A comprehensive benchmark of deepfake detection*. arXiv.<https://arxiv.org/abs/2307.01426>

12. Karkkainen, K., & Joo, J. (2021). *FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation*. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1548–1558).

13. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). *DeepFakes and beyond: A survey of face manipulation and fake detection*. arXiv.<https://arxiv.org/abs/2001.00179>

14. Lin, L., Santosh, Wu, M., Wang, X., & Hu, S. (2025). *AI-Face: A million-scale demographically annotated AI-generated face dataset and fairness benchmark* [Dataset and computer software]. arXiv. [https://arxiv.org/abs/2406.00783;](https://arxiv.org/abs/2406.00783) <https://github.com/Purdue-M2/AI-Face-FairnessBench>

15. Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proenca, H., & Fierrez, J. (2020). *GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection*. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 1038–1048.<https://doi.org/10.1109/JSTSP.2020.3007250>

16. Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. (2021). Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). arXiv. <https://arxiv.org/abs/2007.08457>; https://github.com/ningyu1991/ArtificialGANFingerprints

17. Serengil, S. I., & Ozpinar, A. (2021). *HyperExtended LightFace: A facial attribute analysis framework*. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICEET53442.2021.9659697