Short version:

**Project Title**RUFake: Online Tool for Detecting AI Generated or Altered Photos

**Names**[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)

**Course**Data Science 210 – Spring 2025, Section 6

**Date**April 2025

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 billion1. 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 now a $500B global industry, causing financial and emotional harm. Romance scams alone affect 53% of men and 47% of women. Users lack tools to reliably detect fake images while maintaining privacy. RUFake addresses this by delivering fast, ethical, and private deepfake detection.

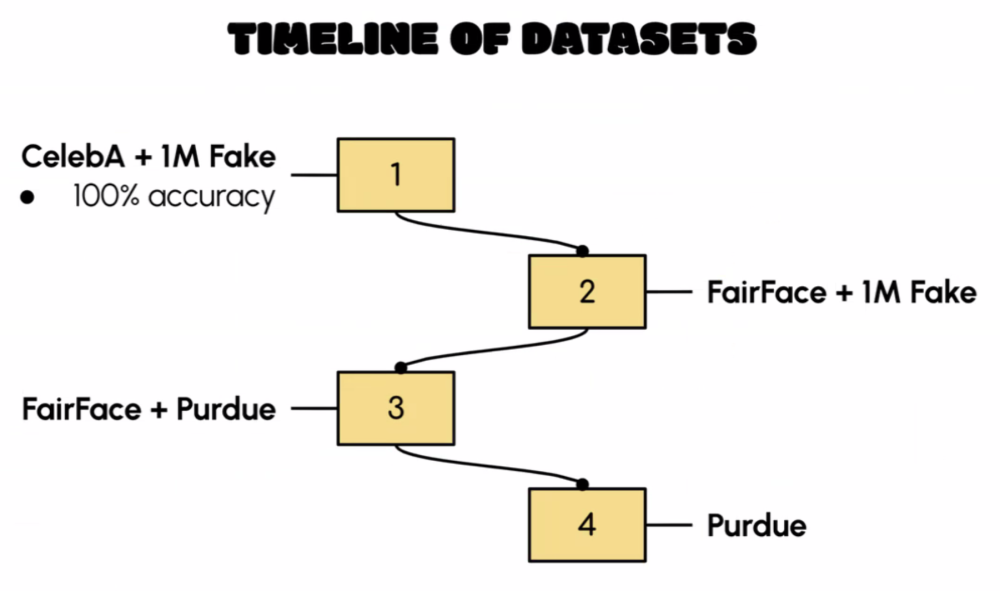
## Privacy and Ethics (By Design)

* 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 altered 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).
  + Mitigate against ethical concerns.
* Our Solution thus minimizes 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.
  + Each prediction includes a confidence score to improve transparency and help users make informed decisions.
  + 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 Source & Data Science Approach  
Our project involved a comprehensive search across more than 15 datasets, five of which we processed extensively. The goal was to construct a fair and representative dataset of real and fake faces. In the early stages, we used DeepFace to annotate gender and ethnicity attributes. Our initial dataset contained 2,000 images—adequate for early experimentation but insufficient for training a robust model.

We then explored several larger datasets. CelebA offered over 200,000 images, but with only 10,000 unique individuals. The 1M AI Generated Faces dataset was based solely on a single GAN method, limiting its diversity in generation methods. FairFace came pre-annotated and was more balanced, but our models achieved 100% accuracy on it too, similar to the challenges we had with the previous datasets.

Eventually, we turned to Purdue’s AI-Face Fairness Bench dataset, which includes both real and fake images generated by 37 different methods and was annotated for gender and ethnicity. Although it took time to obtain permission to use the dataset, it provided the solution we needed. However, the dataset's large size—over 300GB—significantly slowed down our processing pipeline.



Our final dataset included approximately 67,000 images and was carefully balanced by gender and race:

* 39,436 real || 27,914 fake
* 33,621 female || 33,729 male
* 34,171 White || 33,179 non-white (binarized for analysis)
* Age distribution: 43,777 young || 9,665 middle-aged || 1,262 senior || 12,646 other
* All images were resized to 224 × 224 pixels to fit the ViT\_base\_patch16-224 model input requirements.

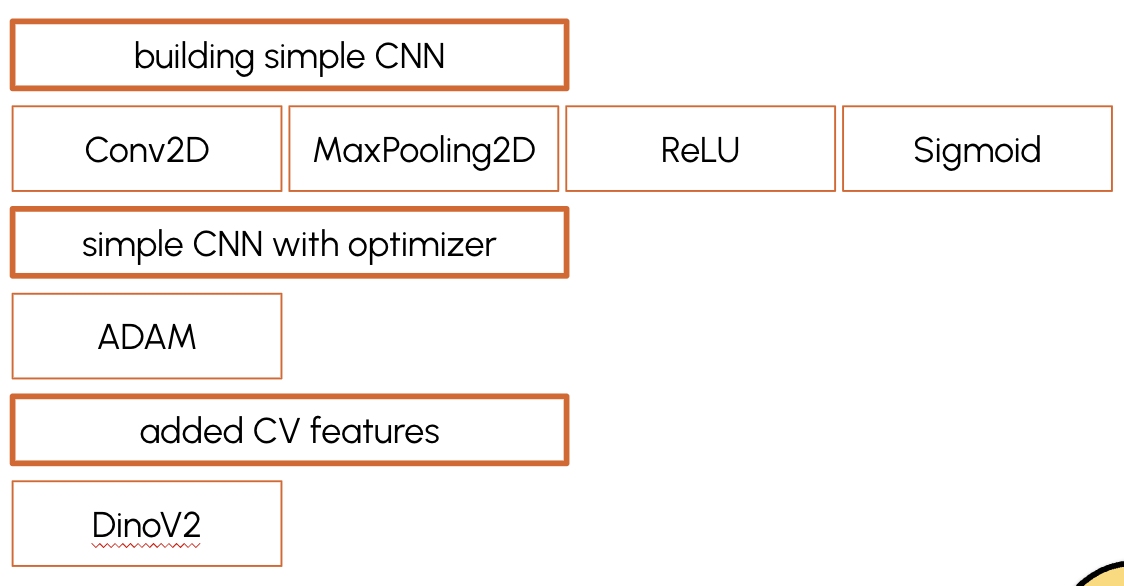
We also experimented with fingerprint-removal methods like GANprintR and synthetic fingerprint injection. While promising, the resulting images were excluded from the final dataset to maintain consistency. Overall, our final dataset was ethically sourced and designed to minimize bias across demographic groups.

Model Design and Development

We began by exploring multiple architectures, including VGG16, ResNet50, DenseNet121, MobileNet, InceptionV3, a simple CNN, and ViT Base Patch16-224. For initial prototyping, we selected the simple CNN paired with the ADAM optimizer due to its lower parameter count, faster training times, and suitability for rapid deployment—especially for mobile integration.

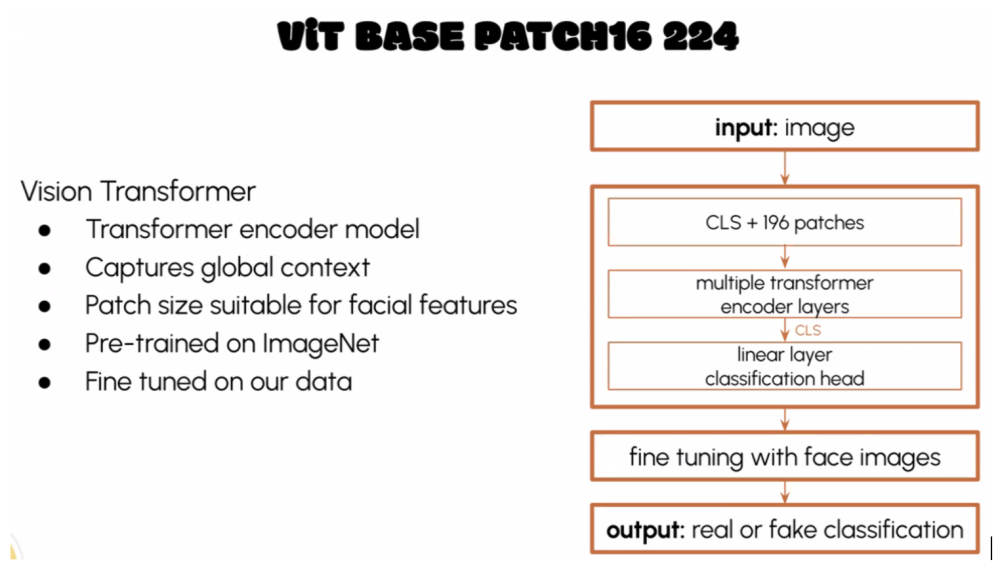
To enhance the CNN's performance, we experimented with a variety of computer vision features: the complex DINOv2, and the simple featuers: Canny edge detection, HOG, Fourier Transform, Error Level Analysis (ELA), edge orientation distribution, color distribution, and dominant colors. Among these, only DINOv2 significantly improved performance, boosting accuracy by 6%, so we retained it in the model.

Our CNN model architecture is shown below.



The second model that we continued to work with was the ViT Base Patch16-224. This vision transformer treats the image as a sequence of patches, allowing it to capture both global context and subtle spatial patterns through self-attention. The patch size was well-suited for detecting facial features. Pretrained on ImageNet, it possessed strong general visual understanding and fine-tuning it on our dataset enabled precise domain adaptation. While the model is not the smallest, it struck a good balance between accuracy and computational efficiency, achieving prediction times of about one second per image.

The model also excels at learning high-dimensional hierarchical features. In the context of fake image detection, it likely identified artifacts such as inconsistent lighting, misaligned facial features, and blurred or overly smoothed textures—common indicators of AI-generated images.



Evaluation

We assessed the CNN and the ViT models using precision, recall, and confusion matrices. We conducted bias audits across age, race, and gender, ensuring balanced performance. Fairness-aware techniques and stakeholder feedback were integrated throughout.

Ultimately, we deployed the —ViT Base Patch16-224—as the final model for our application.

Evaluation of the final model:

Hyperparameters used:

Learning Rate: 2e-5

Loss Function: Cross Entropy Loss

Epochs: 5

Batch size: 32 images

Optimizer: AdamW

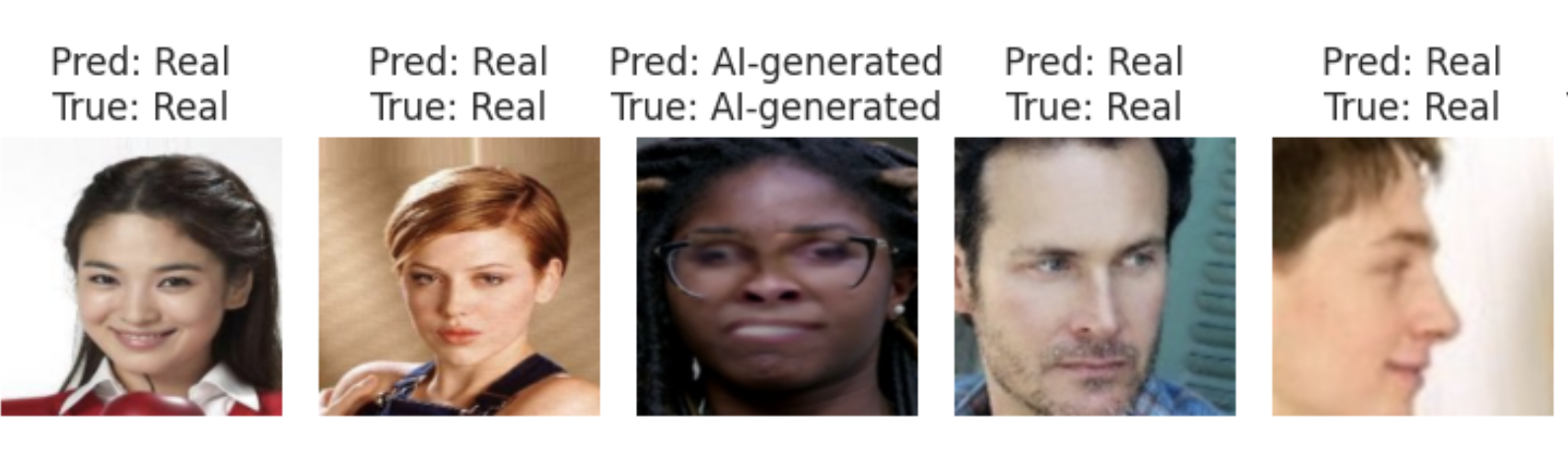
Our model achieved:

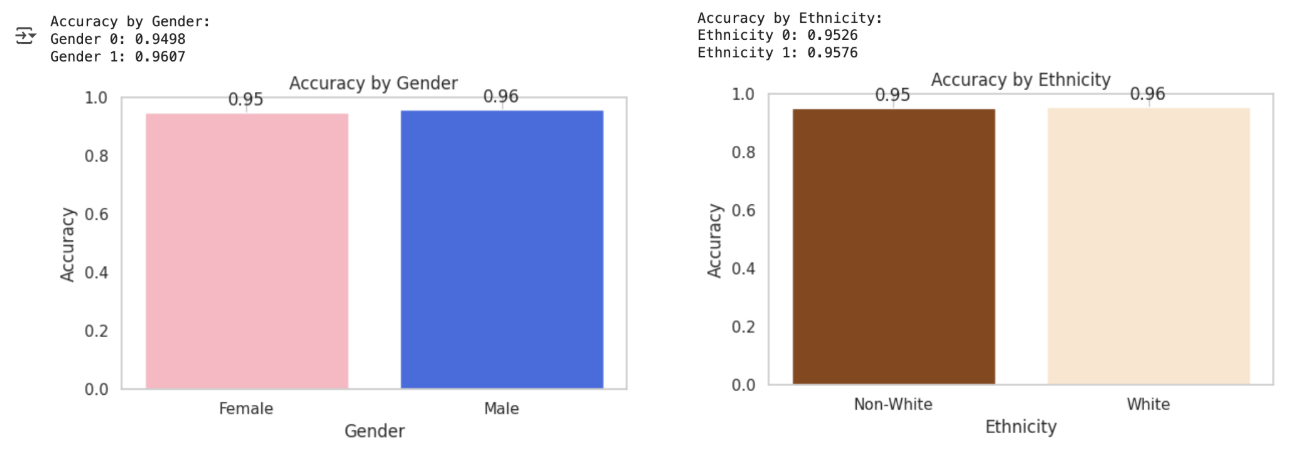
· 95% test accuracy

· 96% overall F1 score

· <1% accuracy difference across demographic groups

· Prediction time: ~1 second/image

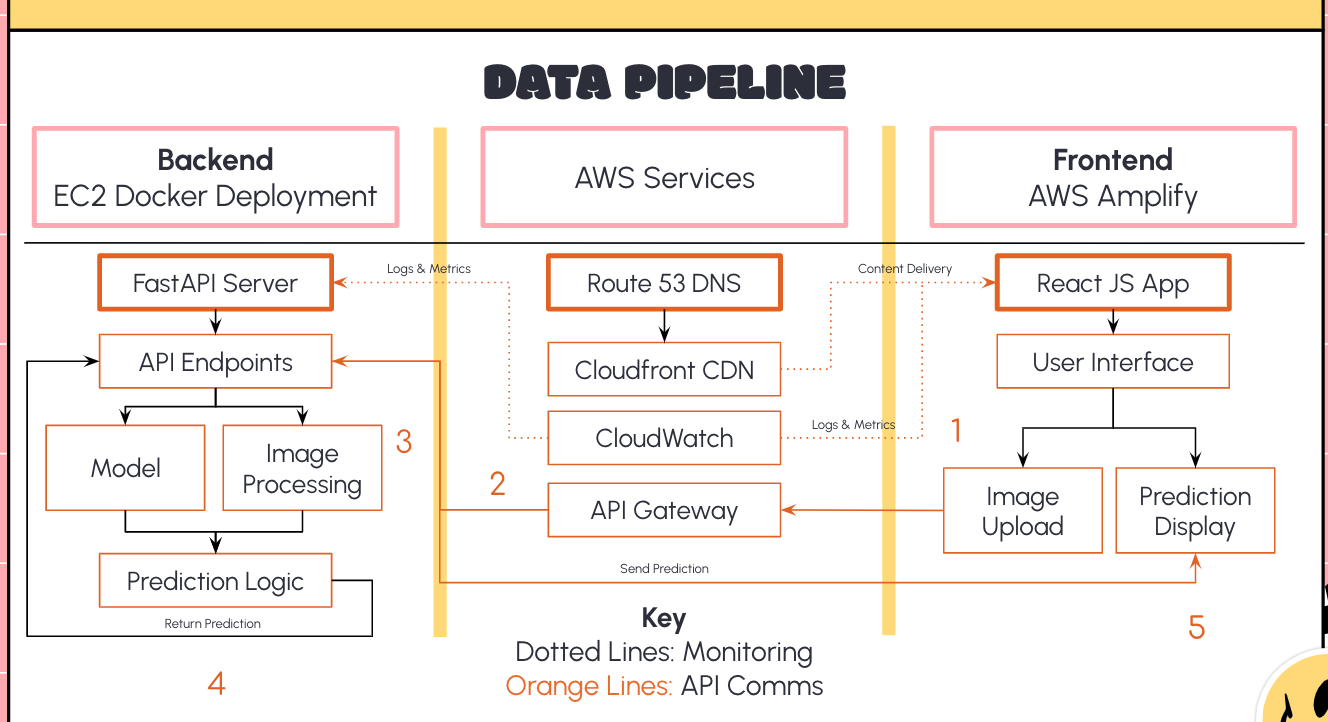




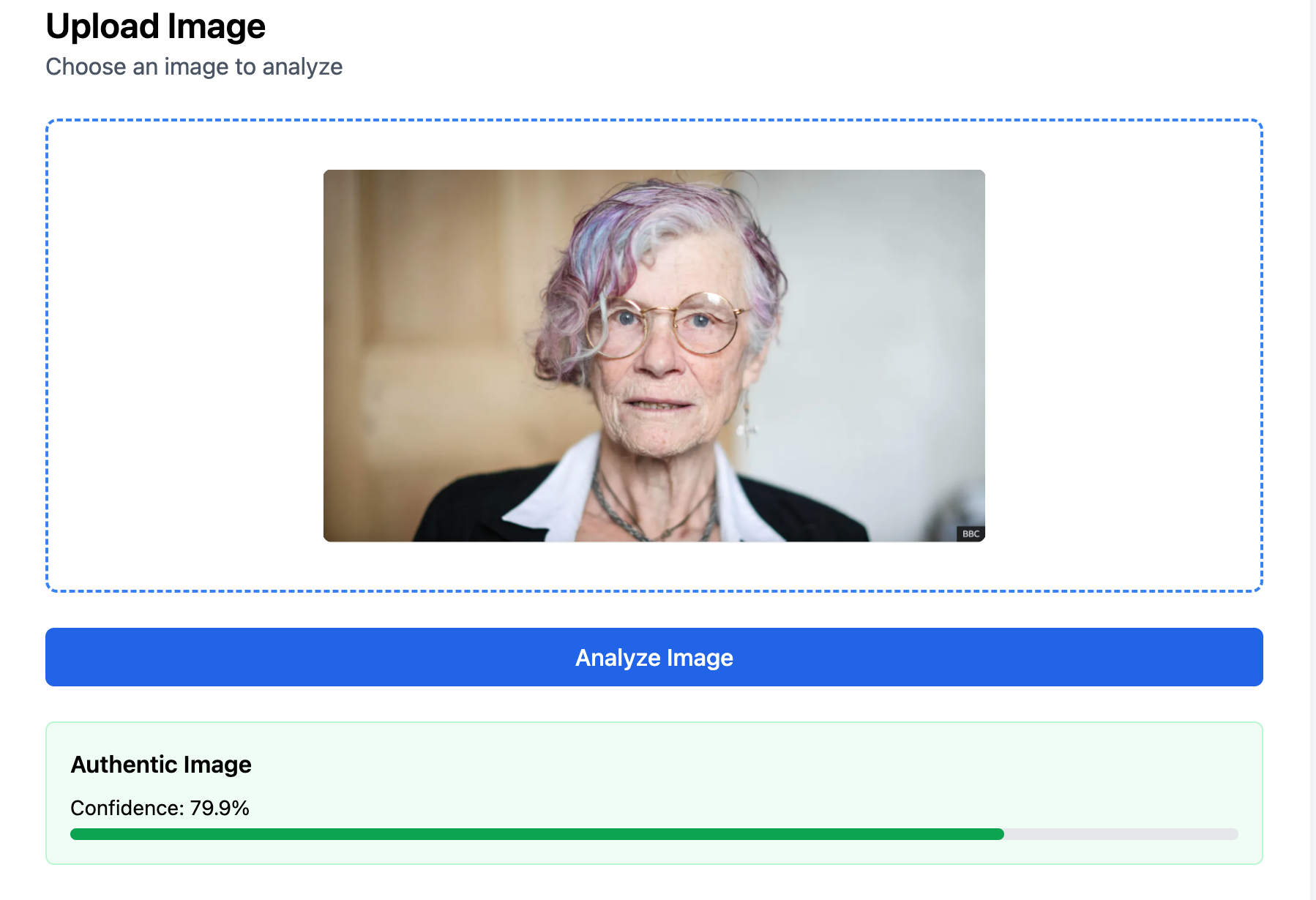
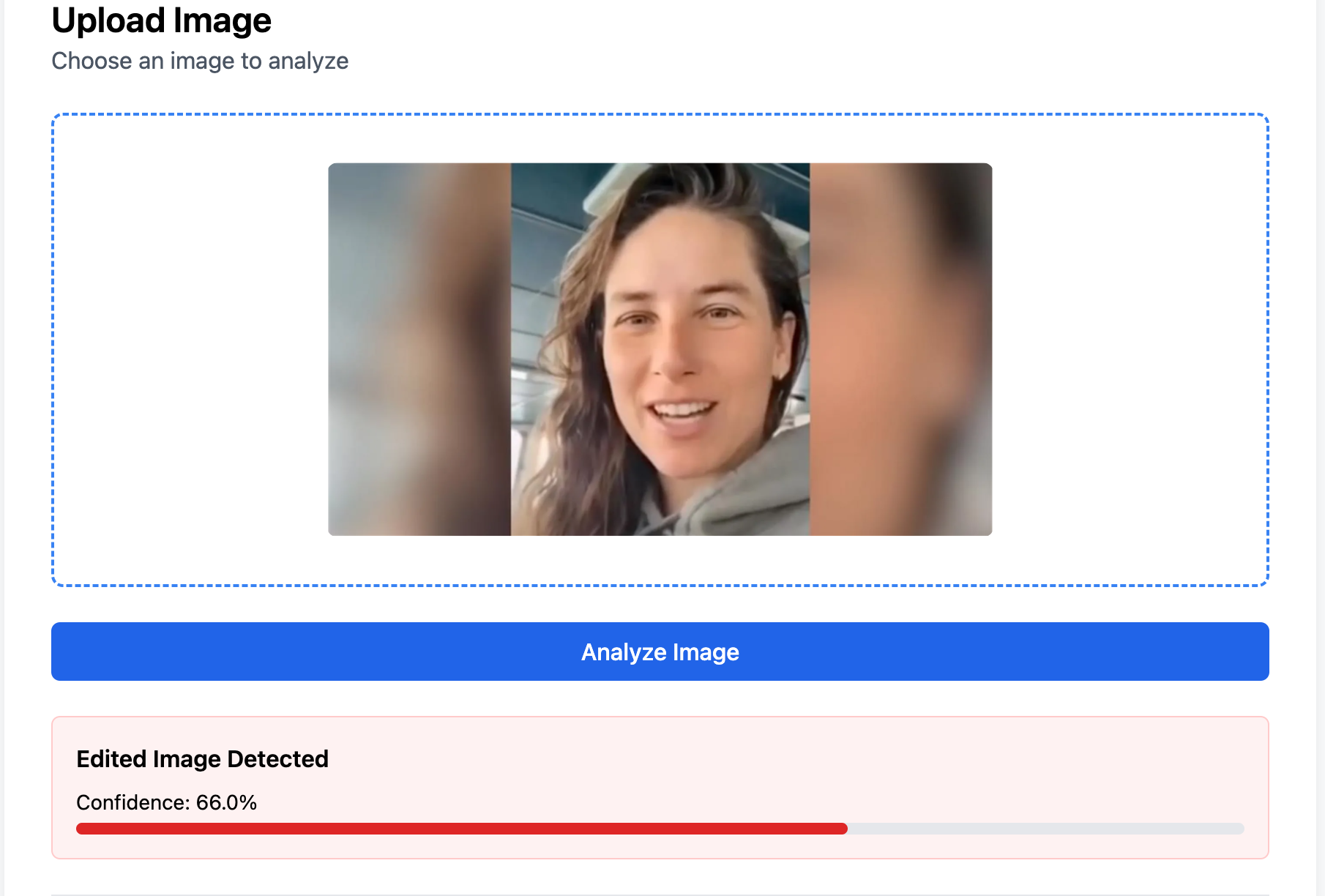
MVP and Backend of the Web Application

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](http://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.



Users will see green if it is likely that an image is authentic, and red if it is likely an altered image. The degree of confidence is designated by a status bar below the prediction. A prediction cannot be below 50%, similar to a softmax, which is our threshold, and high confidence typically reaches around 80%.



Key Learnings & Impact

* We demonstrated that it is possible to distinguish between real and altered images in a robust and fair fashion.
* Finding the right dataset that would enable appropriate training for our model was difficult. We needed sufficient unique images balanced for gender and ethnicity. Combining datasets in many variations 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 (greater than 300GB).
* Balancing for gender and ethnicity in the dataset significantly improved the fair performance of our model compared to previous work done in the field.
* Privacy-preserving design is technically and ethically complex
* Public trust depends on transparency, fairness, and speed

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 altered 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 Puya Vahabi and Prof Todd Holloway
* We are indebted to input from the Berkeley Computer Vision Faculty
* We thank the creators of the datasets and models we were privileged to use.
* We acknowledge feedback from our peers who have shaped the project

References

1. The Economist (2025). Online Scamming.

2. Nissenbaum H. A (2011) Contextual Approach to Privacy Online. *Daedalus*;4:32-48.

3. Wright D (2012). The state of the art in privacy impact assessment. *Computer law & security review*;28:54-61.

4. Allen KC, Davis A, Krishnamurti T (2023). Indirect Identification of Perinatal Psychosocial Risks from Natural Language. *IEEE Trans Affect Comput*.;14:1506-1519. doi: 10.1109/TAFFC.2021.3079282

5. Dankar iK, El Emam K (2013) Practicing differential privacy in health care:. *Trans 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/Poster presented at: Applications of Digital Image Processing XLVII;

7. Whong C (2014). FOILing NYC’s Taxi Trip Data. <https://chriswhong.com/open-data/foil_nyc_taxi/>.

8. Narayan SM, Kohli N, Martin MM (2025). Addressing Contemporary Threats in Anonymised Healthcare Data Using Privacy Engineering. *NPJ Digital Medicine*. *6;8(1):145. doi: 10.1038/s41746-025-01520-6.*

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

**Course**

[**Data Science 210. Capstone**](https://www.ischool.berkeley.edu/courses/datasci/210), Spring 2025

[Class Project Gallery](https://www.ischool.berkeley.edu/courses/project-gallery/284/1/2025)

**Teaser Image**