Supervisor
Lect. Dr. Ciuciu Ioana

Author
Goia Rares-Dan-Tiago

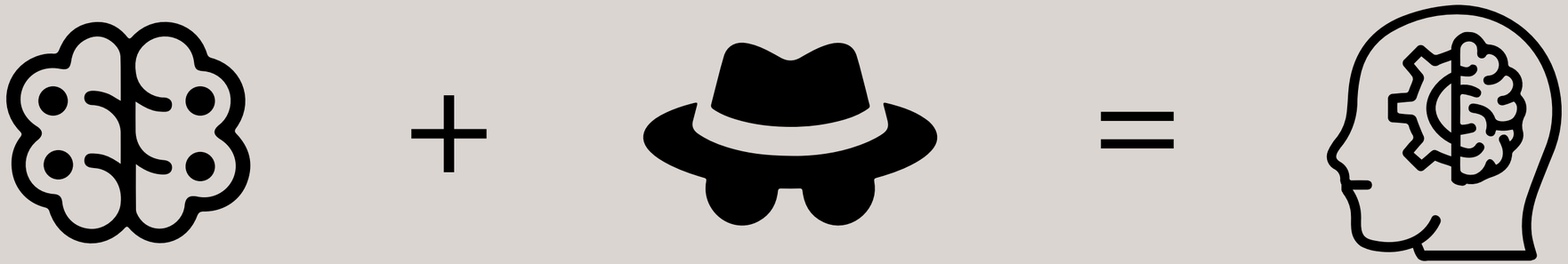
Deepfake Detection in Facial Images

Contents

01
Introduction

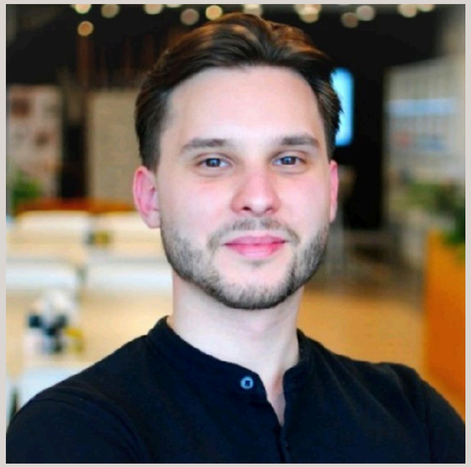
Context

“Deep Learning” + “Faking” = Deepfake



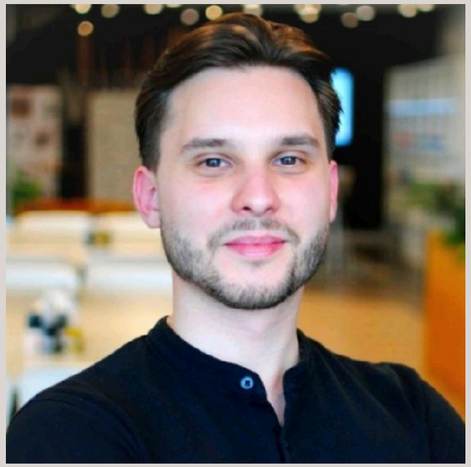
AI-generated synthetic media, like images, videos or audio

Real or Generated?



Real or Generated?

Fake



Real

Fake

Fake

Motivation

Humans struggle with detecting them

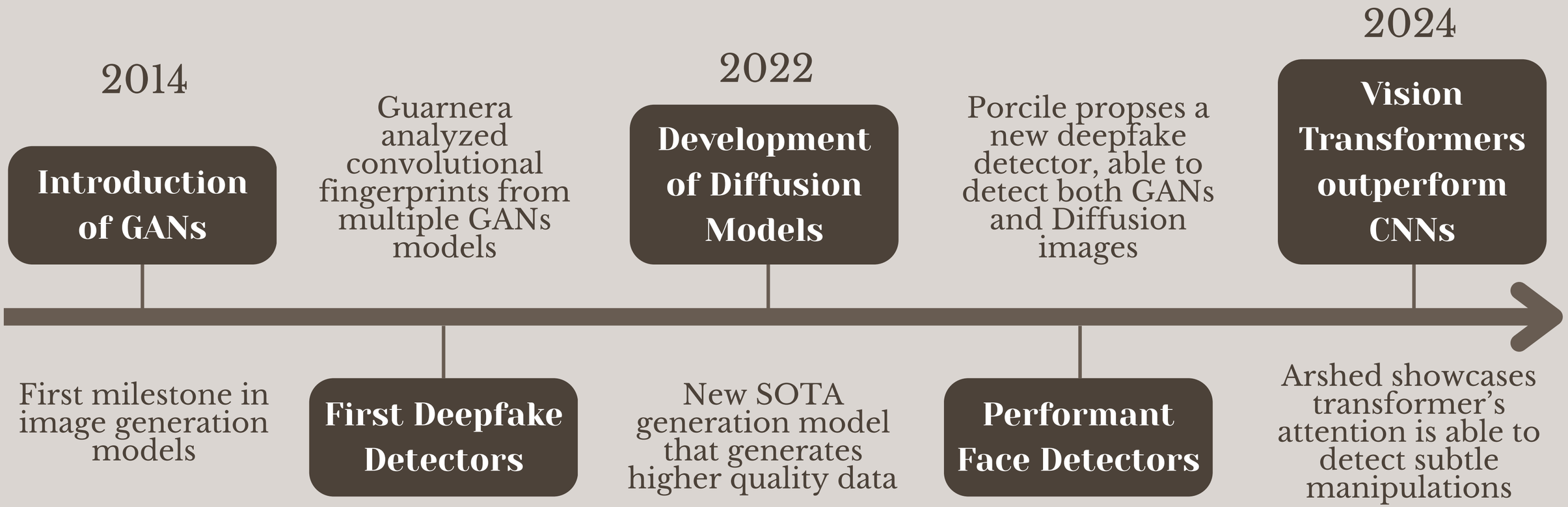
Produce harm by misinformation

Are becoming more and more realistic

No accessible solutions

Public people get impersonated

Related Work



02
Research Objectives

Proposed Solution

Gather and collect Real and AI-Generated images into a dataset. Train and test SOTA models using the new own dataset

Ingrate the models into a user-friendly web application that provides accessibility and explainability

Develop my own model, with a custom architecture and test its performance against the existing models

03
Deepfake Detection

Dataset Construction

Type	Source	No. Images Total	No. Images Used
Real	CelebA Dataset	200,000	4,000
StyleGan1	Collected from documentation	1,000	1,000
StyleGan2	Collected from documentation	1,000	1,000
ThisPersonDoesNotExist	Kaggle Datasets	10,000	1,000
Stable Diffusion XL	Generated by myself	1,000	1,000

Binary Classification

This was made just as a initial step into model creation, to validate that the gathered data can be useful for the models to extract meaningful insights.

EfficientNet-B1
97%

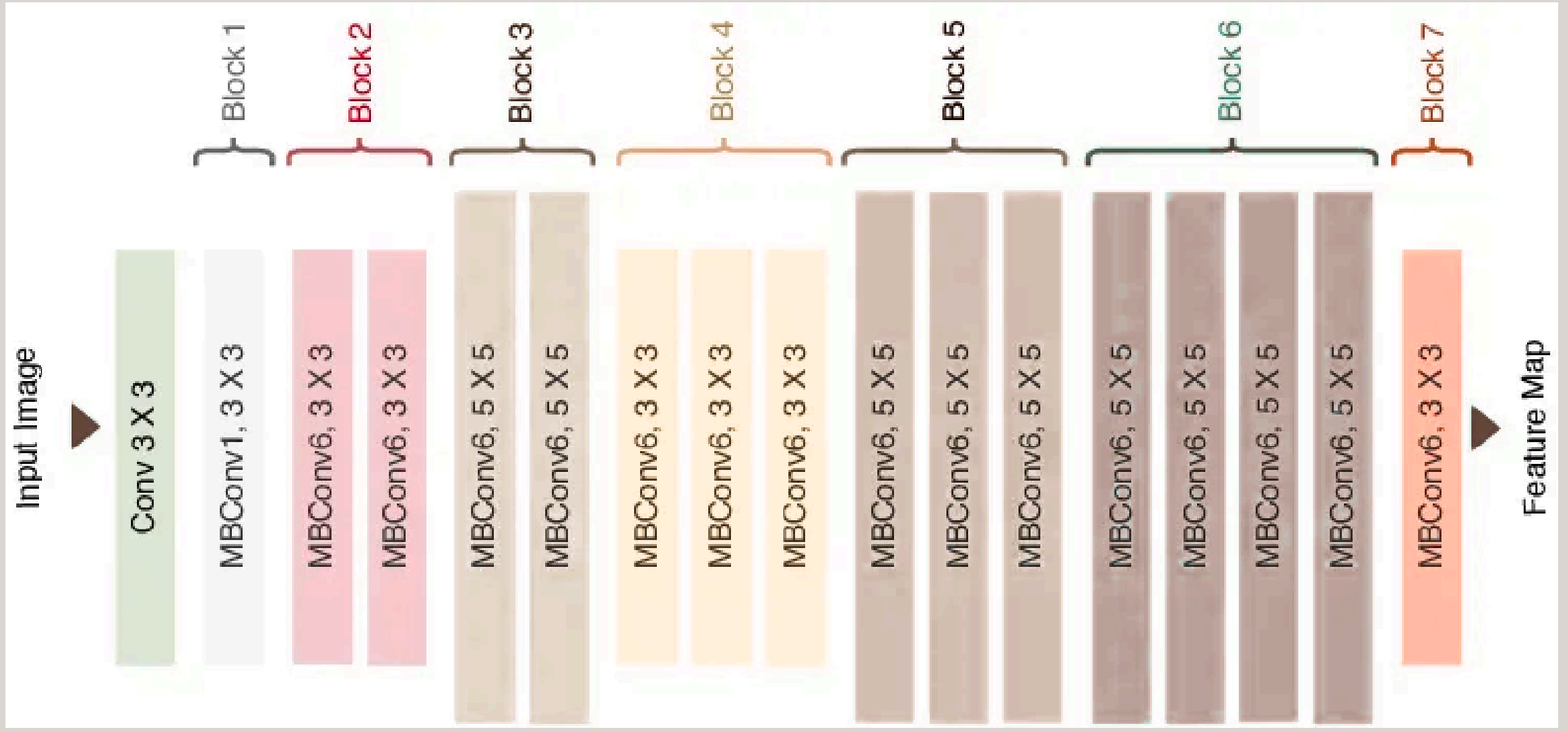
ResNet-50
95%

XceptionNet
88%

These models were chosen as per Porcile's paper. Each model has it's associated F1 Score.

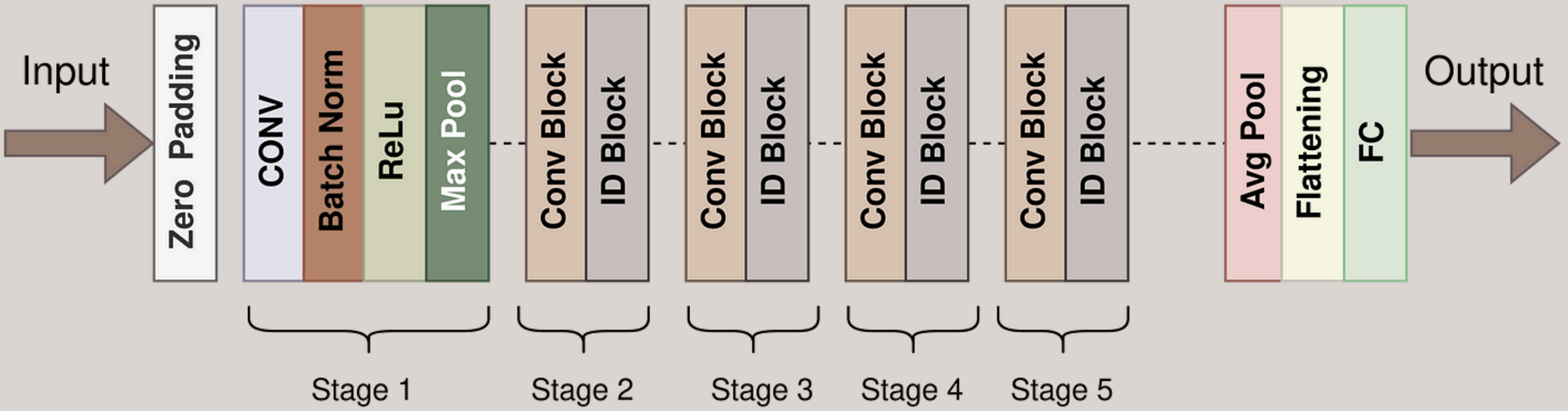
Multiclass Classification

EfficientNet



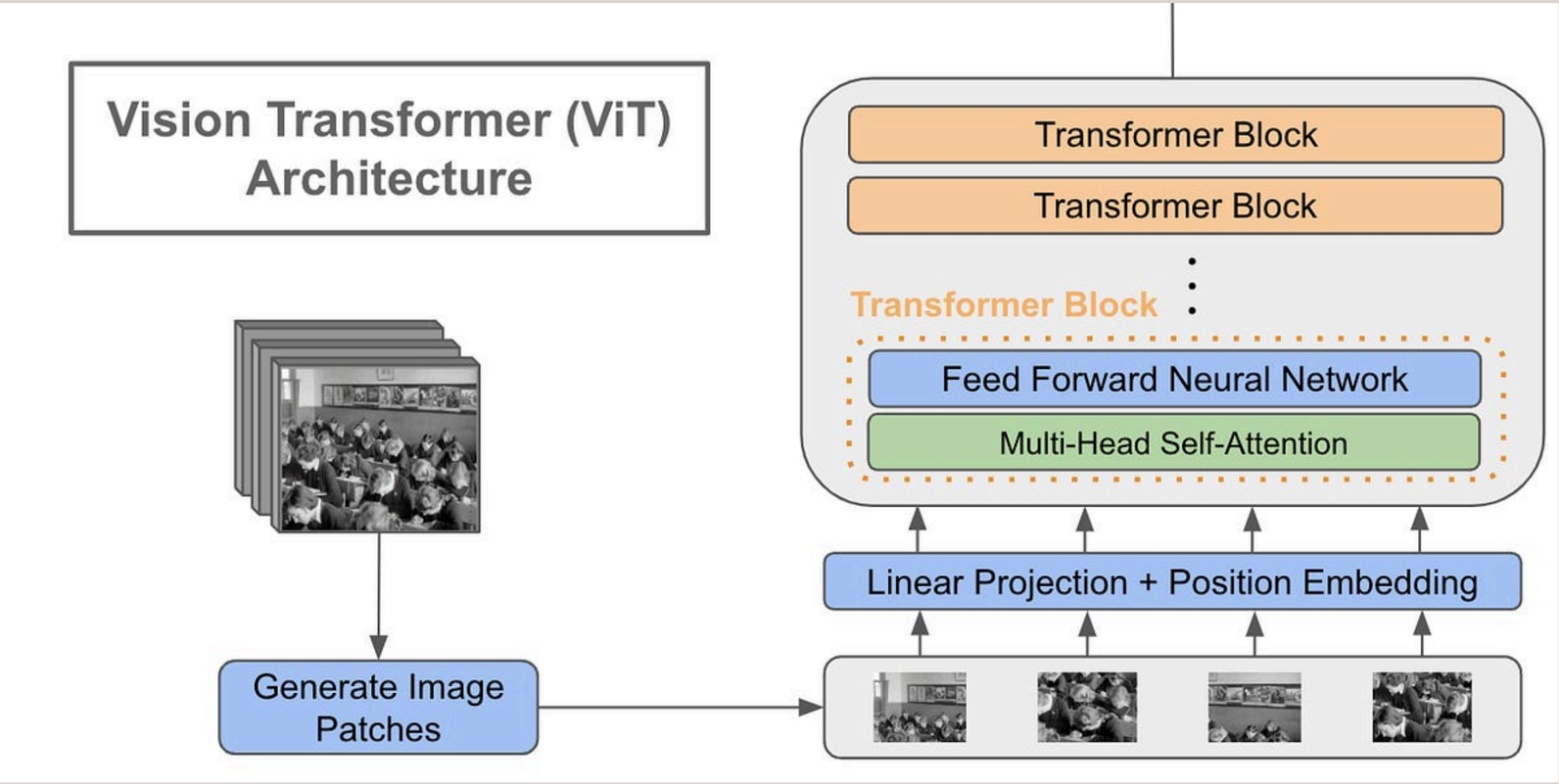
Multiclass Classification

ResNet50



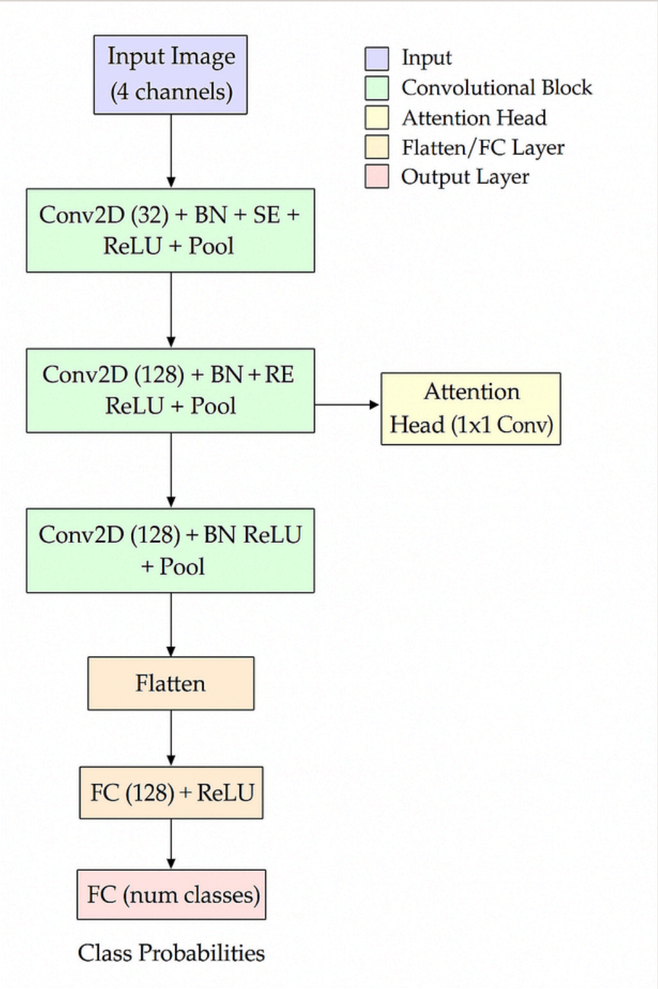
Multiclass Classification

Vision Transformer



Multiclass Classification

Custom CNN



Multiclass Classification

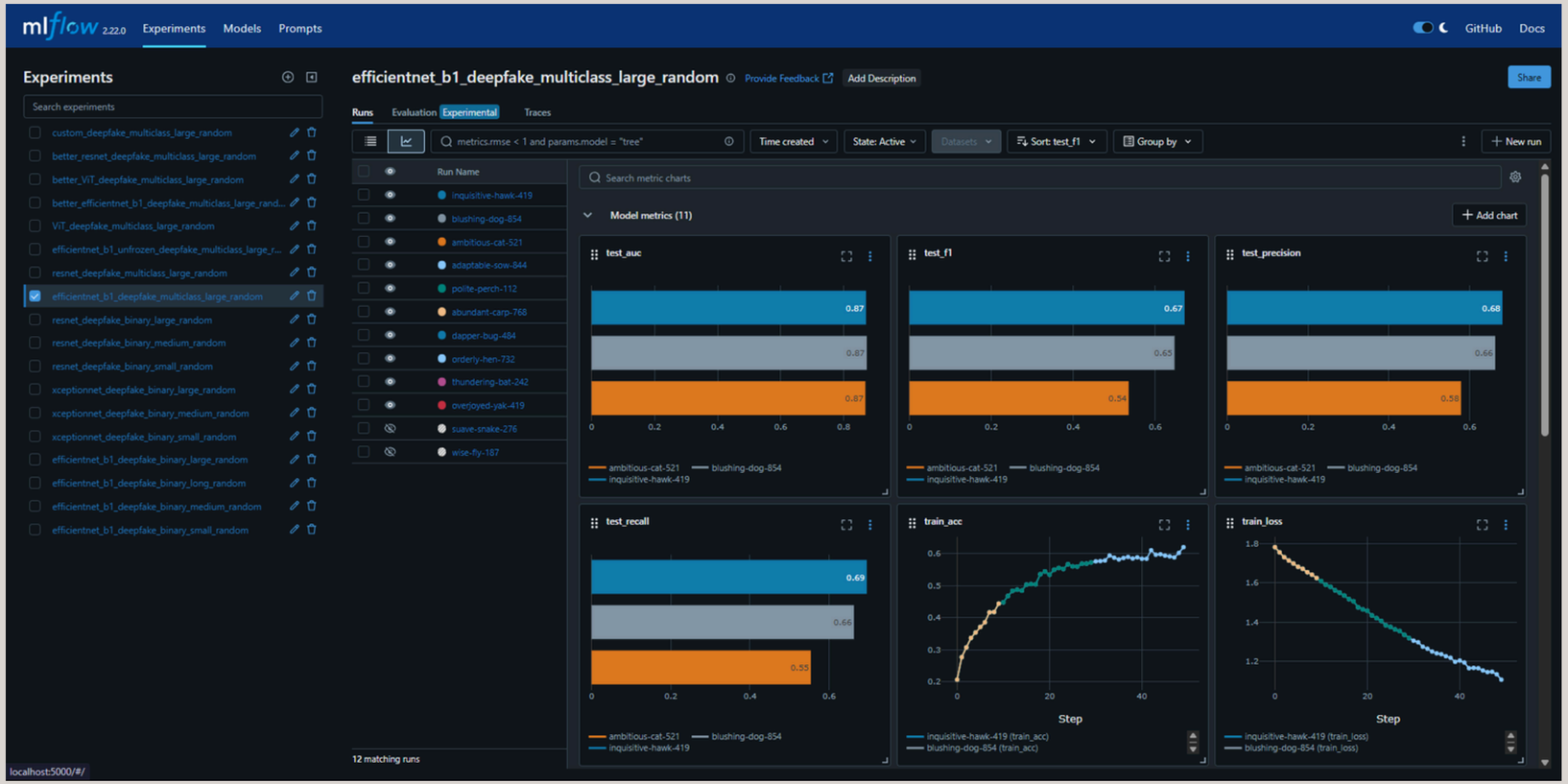
Results

Model	Accuracy	F1-Score
EfficientNet (Mine)	87%	93%
ResNet (Mine)	86%	88%
Vision Transformer (Mine)	91%	92%
Custom CNN (Mine)	80%	82%
EfficientNet (Porcile)	88%	-
Vision Transformer (Arshed)	99%	99%

04
Experiments

MLFlow Tracking

For ease of tracking model's performances, hyperparameters and obtained metrics across all experiments



Hyperparameters

Batch Size: 32 | Epochs: 10–15 | Learning Rate: 1e–4 | Image Height: 218 | Image Width: 178

70% Train | 15% Test | 15% Validation

Binary Classification

Optimizer: AdaGrad

Loss Function: BinaryCrossEntropy

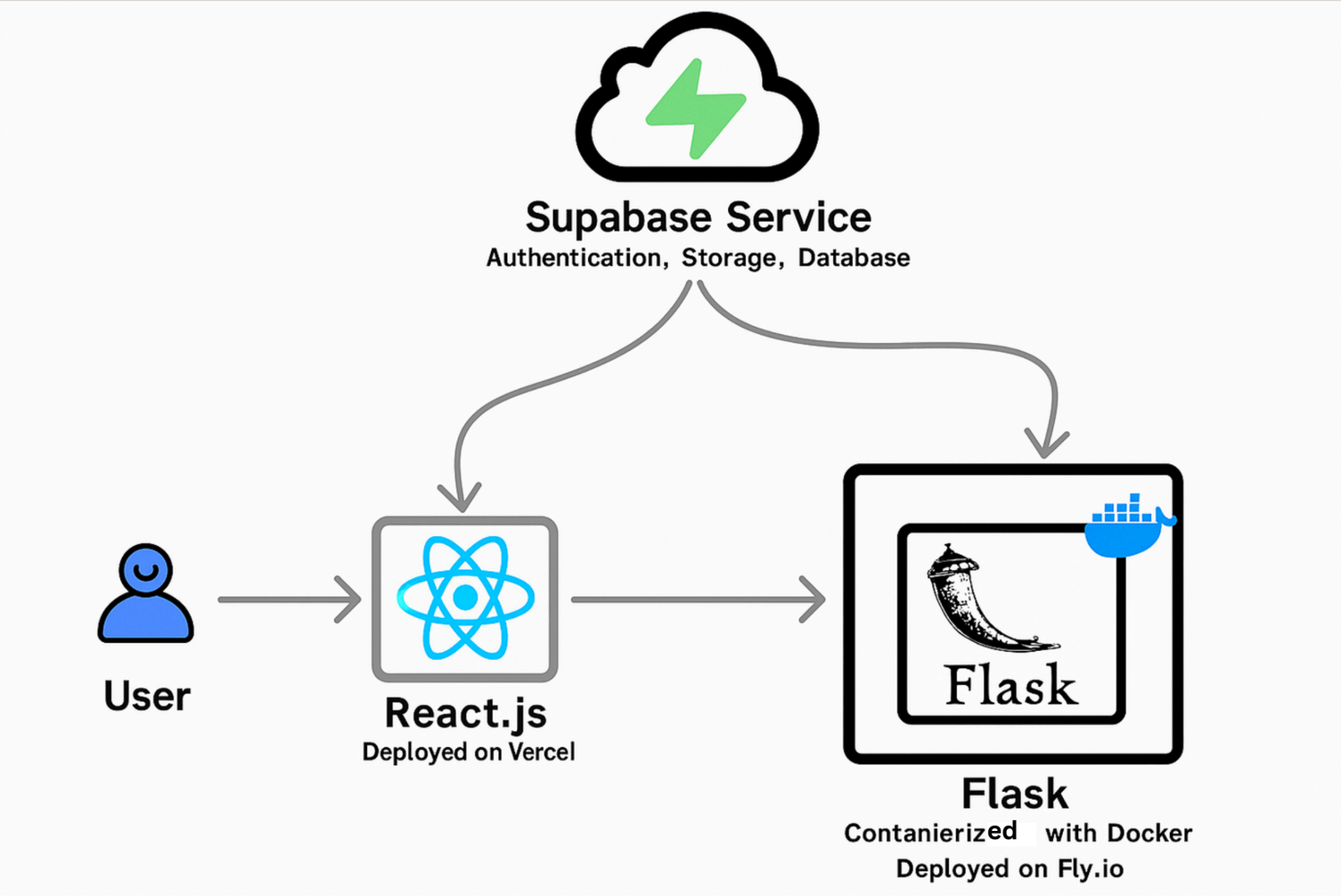
Multiclass Classification

Optimizer: Adam

Loss Function: CrossEntropy

05
Application Development

System Design



Rest API

Detection

GET
/api/models

POST
/api/validate-face

POST
/api/detect

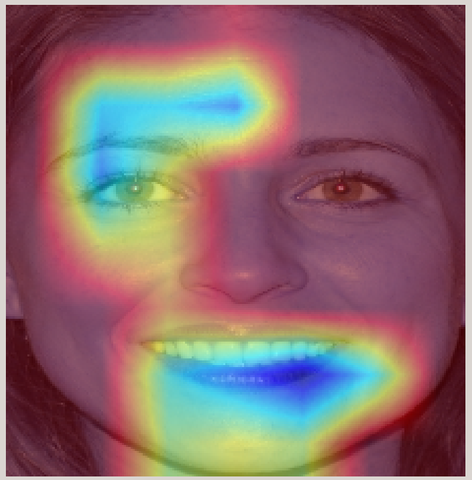
History

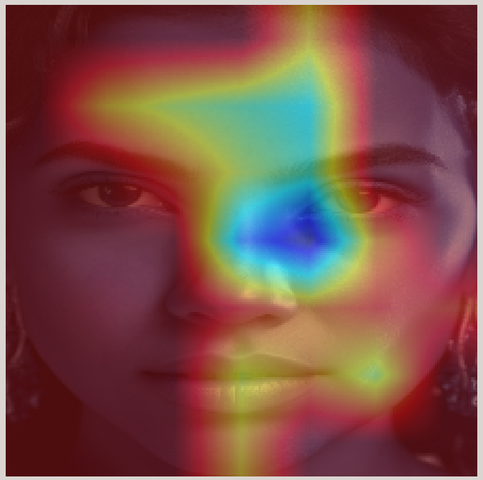
GET
/api/history/images

POST
/api/history/save

DELETE
/api/history/image/id

Explainability: GradCam

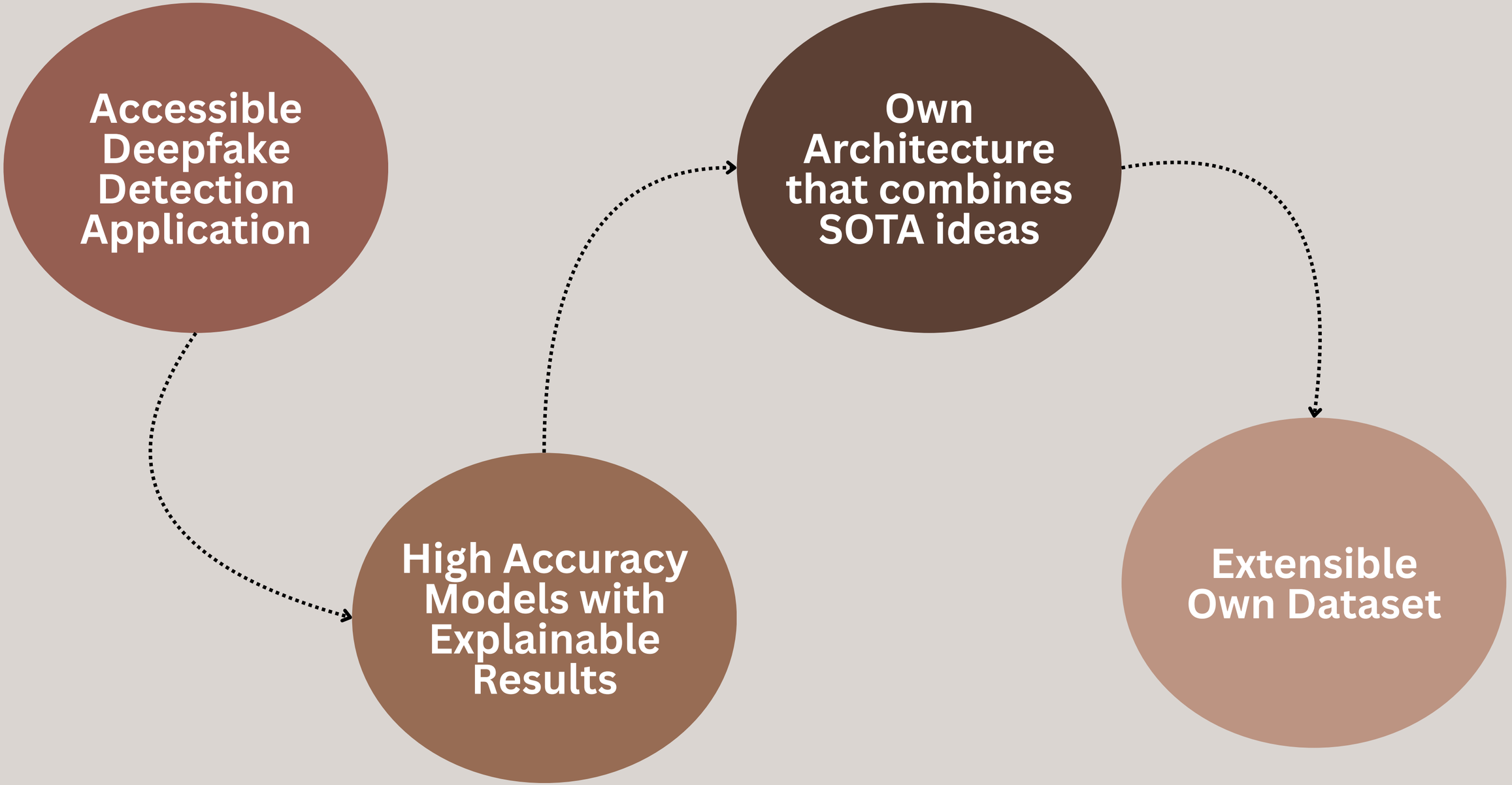





GradCam creates a heatmap that allows the user to understand where the model looked when it took the decision to label an image.

DEMO

06
Conclusions


Accessible Deepfake Detection Application
Own Architecture that combines SOTA ideas
High Accuracy Models with Explainable Results
Extensible Own Dataset

07
Future Work

Future Work

- Extend the research to video-based DeepFakes
- Expand the dataset to capture even more diverse models
- Improve the model’s ability to detect laundering techniques
- Make the app even more accessible with mobile app or chrome extension

08
References

References

1. Finding AI-generated (deepfake) faces in the wild
2. Testing Human Ability to Detect Deepfake Images of Human Faces
3. Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model
4. DeepFake Detection for Human Face Images and Videos: A Survey
5. DeepFake Detection by Analyzing Convolutional Traces
6. New Approaches For Detecting AI-Generated Profile Photos
7. Faster Than Lies: Real-time Deepfake Detection using Binary Neural Networks
8. On the Detection of Digital Face Manipulation
9. Detecting facial image forgeries with transfer learning techniques
10. GenFace: A Large-Scale Fine-Grained Face Forgery Benchmark and Cross Appearance-Edge Learning
11. DeepFeatureX Net: Deep Features eXtractors based Network for discriminating synthetic from real images
12. What is EfficientNet? The Ultimate Guide.
13. The Annotated ResNet-50
14. Vision Transformers ... is using them actually worth it?

Thank You