Image-Based Model Drift Detection

using model inversion and
membership inference attacks

Concurrent Technologies Corporation

Jupyter Notebook Link:
https://colab.research.google.com/drive/1AYklT_aDmMpNuHrKEYgtVfPKXPyoJxIo?usp=sharing

24 Aug 2023
CTC IRAD ML Lab

Outline

Click any header to navigate to the slide

Concurrent Technologies Corporation

What is Model Drift?

Model Drift occurs when the performance of a machine learning model worsens over time

Data Drift : occurs when the data changes from original dataset

e.g., a model will be provided with entirely new data that it has not yet seen.

Concept Drift : occurs when the model's awareness of a certain feature changes

e.g., the data given to the model will be perturbed or manipulated in some way.

images from https://www.flaticon.com/free-icons

Why is this important?

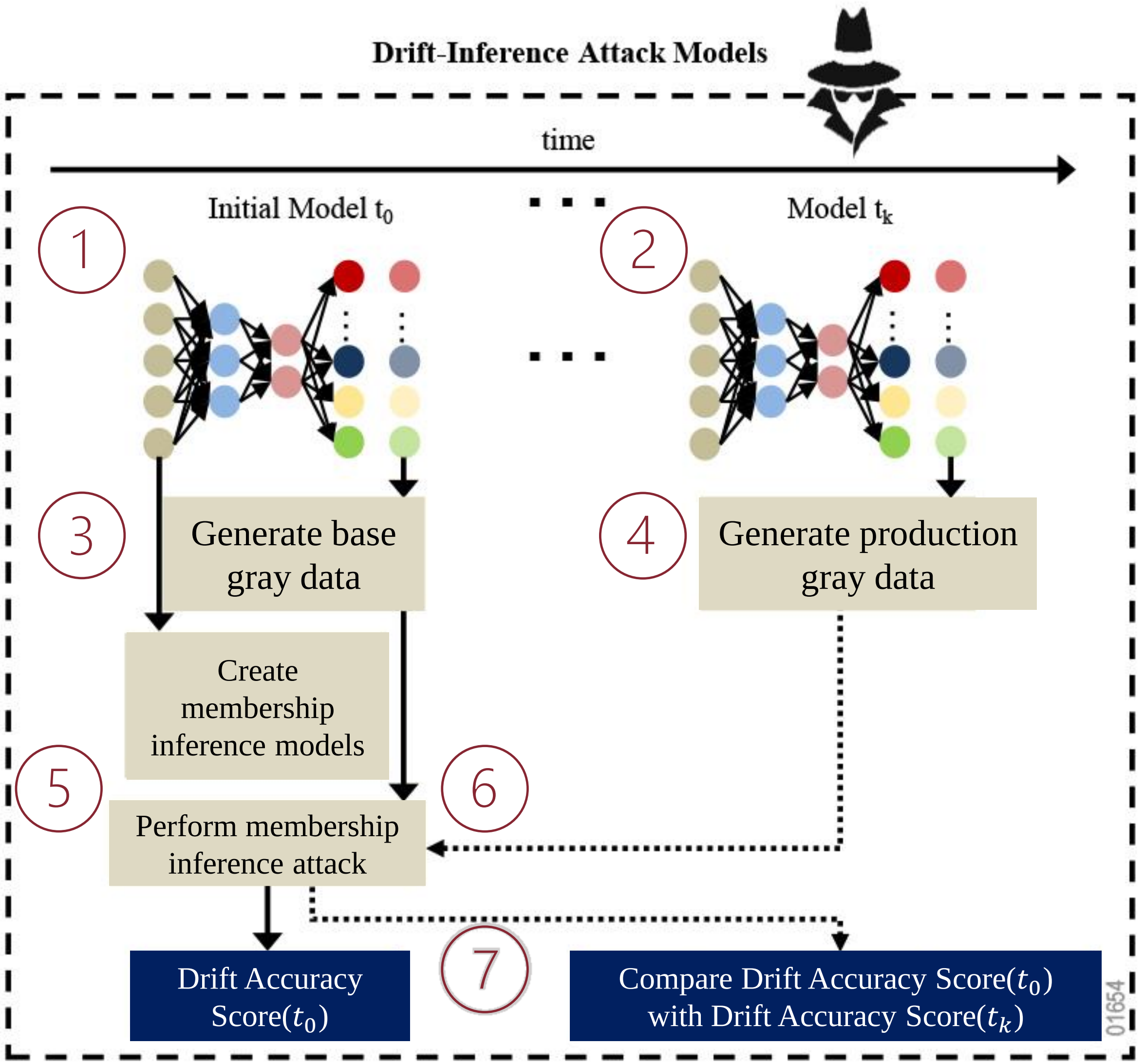
New information is a constant
in real-world applications

Knowing how a model fails
leads to better solutions

Ensures a model's
performance long-term

images from https://mymodernmet.com/navid-baraty-nyc-intersections
and https://www.bing.com/images

Our approach



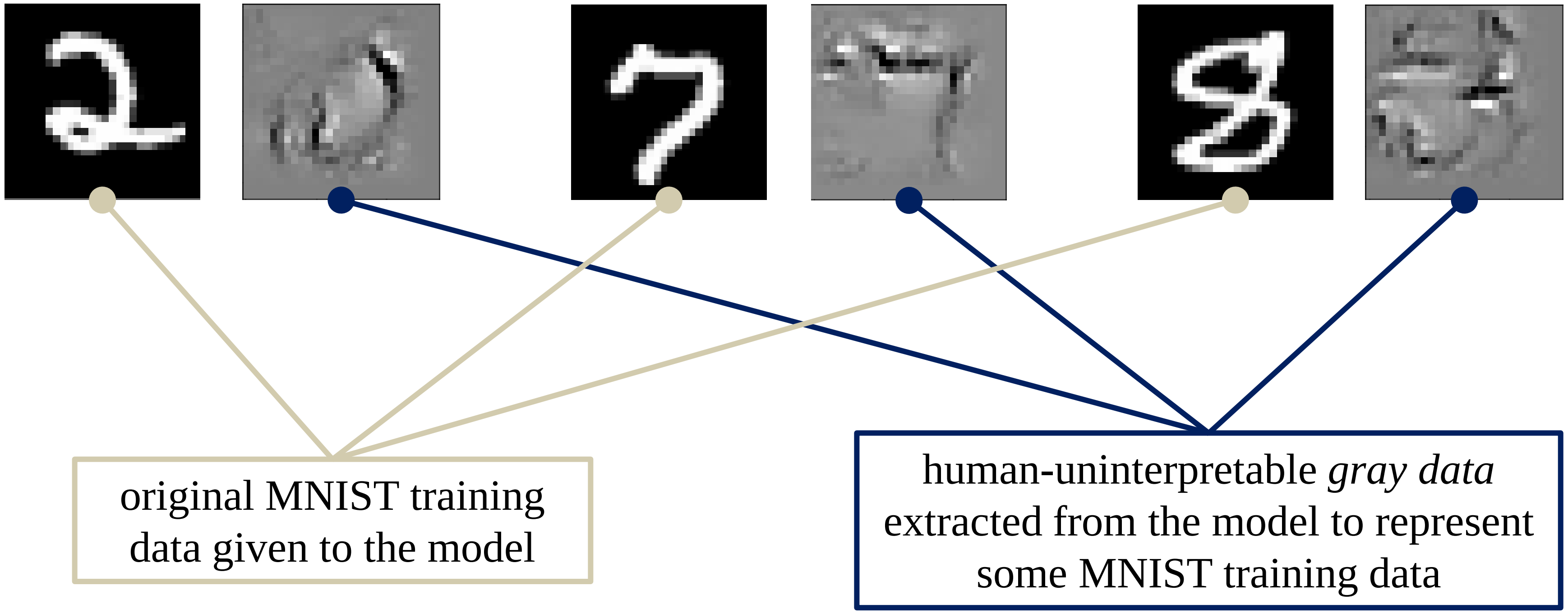
Detecting ***concept drift*** in image data

7 steps to test the base model's definition of the concepts (or classes) of image datasets:

- **#1-2** : create models on different portions of the dataset
- **#3-4** : create ***gray data*** for each model
- **#5-6** : ***infer membership*** of gray data within base model
- **#7** : check scores to detect drift

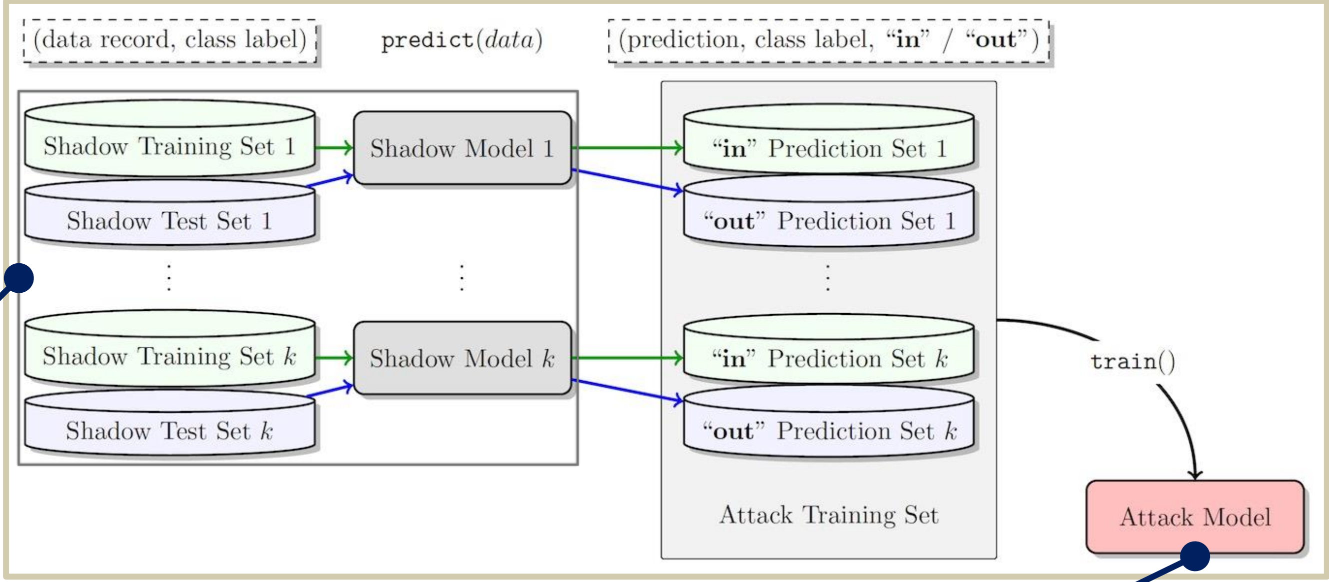
What is Gray Data?

Model inversion attacks use a classifier to attempt to recreate the training data given to a model as "*gray data*"



What is Membership Inference?

Membership inference attacks aim to predict if a specific instance of data was used in the training data for the model

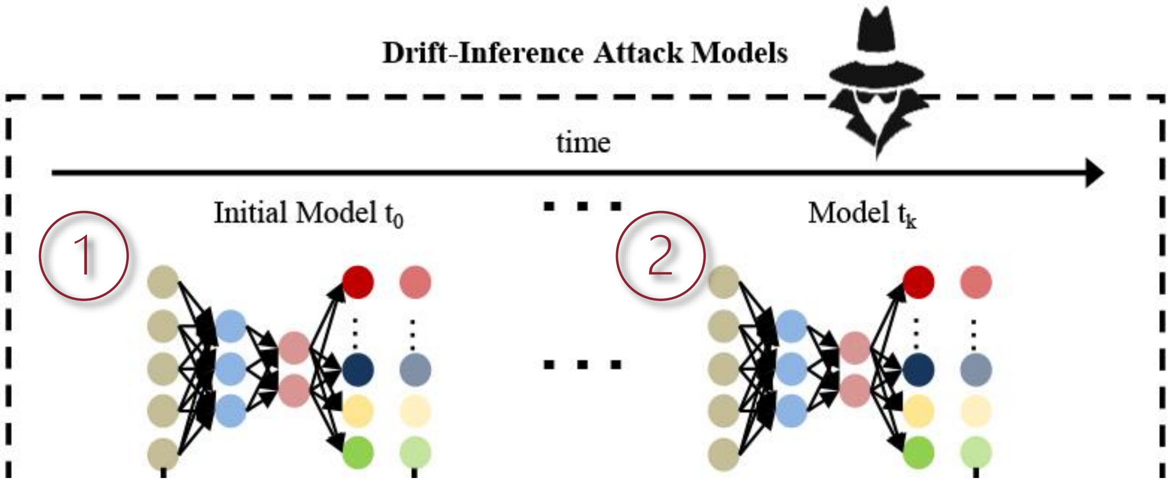


shadow models created by feeding random examples the model to determine key features

final model can predict whether a data point was in the training data of a model

image from https://bdtechtalks.com/2021/04/23/machine-learning-membership-inference-attacks/

Model Creation

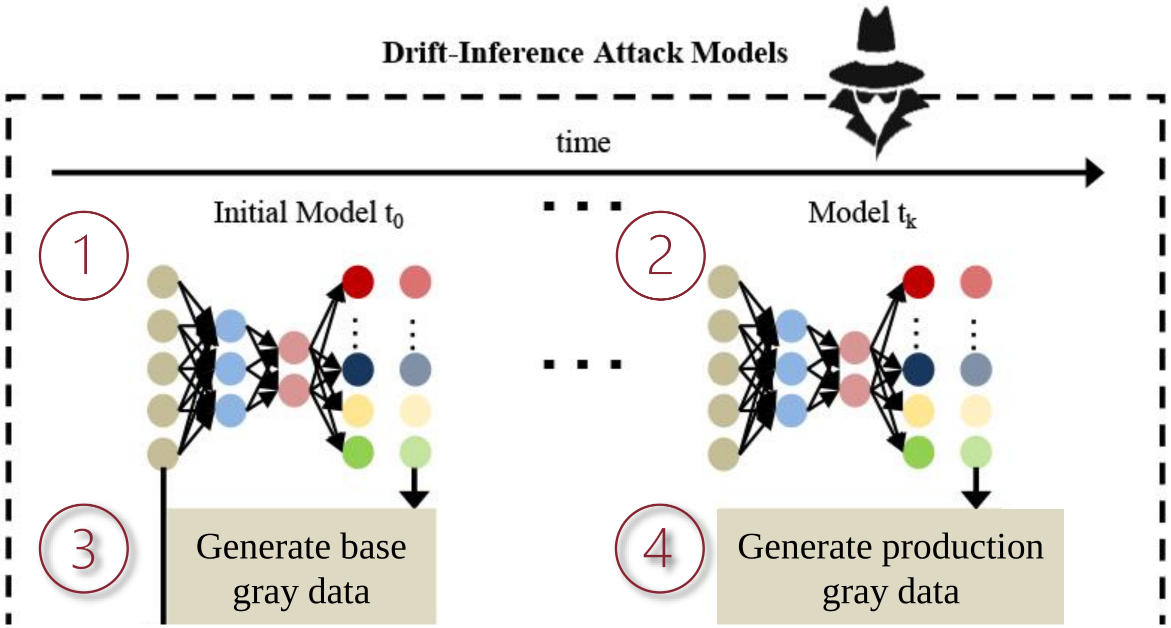


Split training data in original dataset into base data and production data, then…

1 Create **base model** from base data

2 Create **production model** from production data

Gray Data Generation



3 Run a *model inversion* attack on the base model to generate **base gray data**

determines the base model's concept of the classes from the base data

4 Run a *model inversion* attack on the production model to generate **production gray data**

determines the production model's concept of the classes from production data

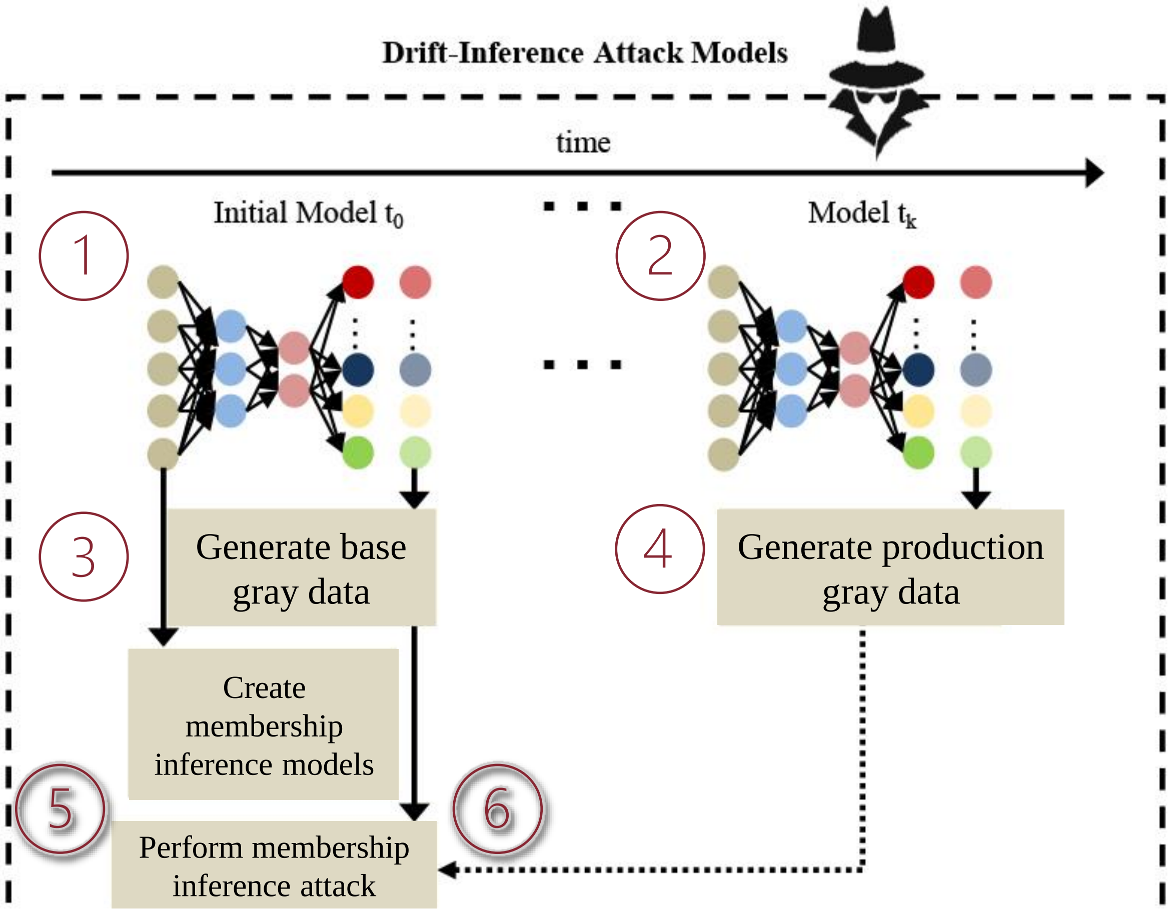
Membership Inference Attacks

5 Run *membership inference* attack on base model with base gray data

base drift score determines whether the concept of a class has drifted

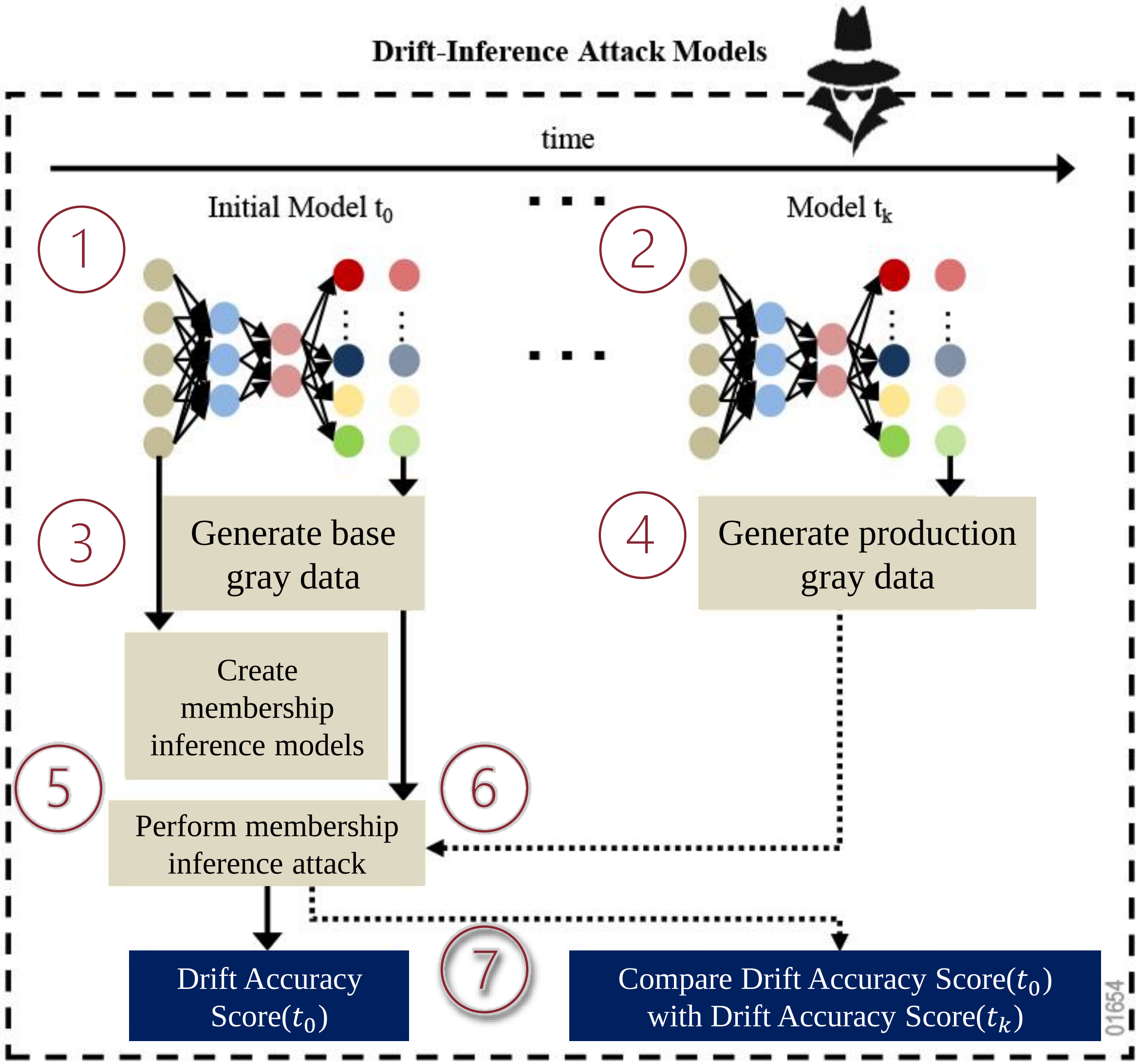
6 Run *membership inference* attack on base model with production gray data

production drift score determines whether the concept of a class has drifted



Model Drift Detection

7 If ***production drift accuracy score*** << ***base drift accuracy score***, then *concept drift confirmed*



Does our approach work?

	base model test accuracy	production model test accuracy
CIFAR-10	0.5738	0.5969
EMNIST	0.8200	0.8246

results collected using an average of three tests from our published Jupyter notebook

no model drift is indicated

Does our approach work?

results using our published Jupyter notebook	Rule-Based Membership Inference Attacks		Black-Box Membership Inference Attacks	
	accuracy with base gray data	accuracy with production gray data	accuracy with base gray data	accuracy with production gray data
CIFAR-10	1.0	0.6666	1.0	0.3611
EMNIST	1.0	0.5352	0.9516	0.1713

results collected using an average of three tests from our published Jupyter notebook

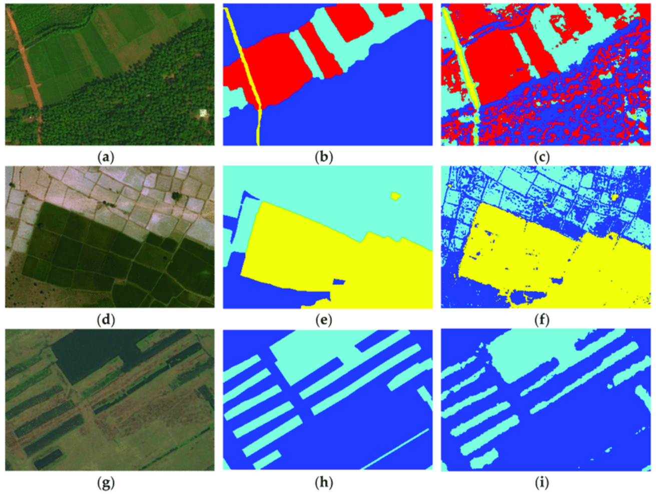
production drift acc << base drift acc

model drift detected

What is the future direction with this method?

Detection → Mitigation

detect drift in a model and recover from drift with an updated model



Overhead Imagery Datasets

detect drift in a model that operates on overhead imagery

images from https://www.researchgate.net/figure/Segmentation-results-for-satellite-images-from-our-dataset-part-1-23-different_fig2_347844387 and https://www.anolytics.ai/blog/how-to-improve-machine-learning-model-performance/